Resolving controversies about Spatial-Numerical Associations:

Insights from different tasks, cultures, stimuli, and time points

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Lilly Zoe Roth
aus Schwetzingen

Tübingen
2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	10.02.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Hans-Christoph Nürk
2. Berichterstatter:	Prof. Dr. Ulf-Dietrich Reips
3. Berichterstatter:	Assoc. Prof. Dr. Michele Vicovaro

Table of contents

Table of contents 3
List of figures 6
List of tables 7
Abstract (English) 8
Zusammenfassung (German) 10
Abbreviations 12
List of publications 13
1. Introduction 15
1.1. Spatial-Numerical Associations 16
1.2. Theoretical accounts of the SNARC effect 17
1.2.1. The Mental Number Line account 18
1.2.2. The Working Memory account 19
1.2.3. Polarity and markedness accounts 22
1.2.4. Other theoretical accounts 24
1.3. Influences on the SNARC effect 27
1.3.1. The SNARC effect in automatic and intentional processing tasks 27
1.3.2. The SNARC effect in different cultures and languages 30
1.3.3. The SNARC effect in different stimuli 34
1.3.4. The SNARC effect's inter- and intraindividual variability 37
1.4. Internet-based data collection 39
1.5. Open Science practices 41
1.6. Research aims 43
2. Results 49
2.1. Quantification of the SNARC effect 49
2.2. Short summary of the main results in each study 50
2.2.1. Study 1: High SNARC automaticity shown in non-semantic tasks 50
2.2.2. Study 2: Different shapes of the SNARC effect in different semantic tasks 50
2.2.3. Study 3: Influence of culture and script direction on the strength of the SNARC effect 51
2.2.4. Study 4: Coexistence of spatial ordinality representations based on working memory and long-term memory 52
2.2.5. Study 5: Partly fixed and partly flexible nature of the SNARC effect 52

2.2.6. Study 6: High inter- and intraindividual variability of the SNARC effect 53
3. Discussion ... 54
3.1. Insights into influences on the SNARC effect ... 54
3.1.1. The SNARC effect in automatic and intentional processing tasks ... 54
3.1.2. The SNARC effect in different cultures and languages ... 58
3.1.3. The SNARC effect in different stimuli ... 61
3.1.4. The SNARC effect's inter- and intraindividual variability ... 67
3.2. Implications for theoretical accounts of the SNARC effect ... 71
3.2.1. The Mental Number Line account ... 71
3.2.2. The Working Memory account ... 73
3.2.3. Polarity and markedness accounts ... 75
3.2.4. Other theoretical accounts ... 78
3.3. A computational contextualization model for the SNARC and SNARC-like effects ... 79
3.3.1. Aims and basic model architecture ... 80
3.3.2. Preliminary optimization results ... 84
3.3.3. Future model improvements and extensions ... 90
3.3.4. Embedding the model within existing theoretical accounts ... 91
3.4. Limitations and future research directions ... 94
4. Conclusion ... 97
References ... 98
Eidesstattliche Erklärung ... 126
Acknowledgments ... 127
Appendix ... 129
A1. Study 1 ... 129
A1.1. Publication of Study 1 ... 129
A1.2. Corrigendum for Study 1 ... 165
A2. Study 2 ... 168
A2.1. Manuscript for Study 2 ... 168
A3. Study 3 ... 277
A3.1. Publication of Study 3 ... 277
A3.2. Supplementary materials for Study 3 ... 302
A4. Study 4 ... 319
A4.1 Publication of Study 4 ... 319
A4.2. Supplementary materials for Study 4 ... 343
A5. Study 5 ... 357

A5.1. Publication of Study 5 357
A5.2. Supplementary materials for Study 5 383
A6. Study 6 394
A6.1. Publication of Study 6 394
A7. A computational contextualization model 425
A7.1. Details on spatial bias distributions 426
A7.2. Details on data simulation 427
A7.3. Details on parameter optimization 429
A7.4. Detailed preliminary optimization results 433
A7.5. Comparison of observed data and simulated data 440

List of figures

Figure 1-1	Graphical overview of the focus of investigation in each study	44
Figure 1-2	Graphical overview of the tasks used in each study	45
Figure 3-1	Mean RT per stimulus in go trials for all four memorized sequences used in Experiment 3 in Study 4	66
Figure 3-2	Spatial bias distributions in contexts that may be activated in number tasks	82
Figure 3-3	Spatial bias distributions in contexts that may be activated in letter tasks	83
Figure 3-4	Comparison of observed and simulated SNARC effects in three number tasks	89
Figure A7-1	Belief distance computation for stimulus number 2 in an example trial	428
Figure A7-2	Comparison of observed and simulated SNARC effects in three number tasks	441
Figure A7-3	Comparison of observed and simulated LTM-based OPE in three letter tasks	442
Figure A7-4	Comparison of observed and simulated WM-based OPE in two letter tasks	443

List of tables

Table 1-1	Open Science practices applied in each study with links to publicly available materials	42
Table 1-2	Overview of methods and results of all studies included in this dissertation	46
Table 3-1	Optimized weights for all schemata in the computational contextualization model	86
Table A7-1	Default parameters for each task in the computational contextualization model	432
Table A7-2	Mean RMSE across participants per task for default versus optimized parameters in the computational contextualization model	434
Table A7-3	Parameters per task averaged across participants resulting from Optimization 1 of the computational contextualization model	435
Table A7-4	Parameters per task averaged across participants resulting from Optimization 2 of the computational contextualization model	436
Table A7-5	Variability of parameters and RMSE averaged across participants in Optimization 1 in the computational contextualization model	438
Table A7-6	Variability of parameters and RMSE averaged across participants in Optimization 2 in the computational contextualization model	439

Abstract (English)

Numbers are fundamental to everyday life, and understanding their cognitive processing and mental representations is therefore essential. A key phenomenon in the field of numerical cognition is the Spatial-Numerical Association of Response Codes (SNARC) effect, which refers to faster responses to small numbers with the left hand and to large numbers with the right hand (Dehaene et al., 1993). This dissertation addresses several controversies surrounding the SNARC and SNARC-like effects in six large (mostly online) studies. Typically, the SNARC effect is investigated in bimanual decision tasks using number stimuli from 1 to 9 (excluding 5). Task-relevant features include number magnitude or ordinality (e.g., magnitude classification: smaller vs. larger than 5), other semantic features (e.g., parity judgment: odd vs. even), or non-semantic features (e.g., color judgment: blue vs. yellow). We found that the SNARC effect follows a continuous pattern in parity judgment (i.e., gradually increasing right relative to left association with increasing number magnitude), but a more categorical pattern in magnitude classification (i.e., similar left association for small and right association for large numbers; Studies 2 and 3). Despite mixed prior findings, we also observed a SNARC effect in color judgment, albeit weaker and driven only by small numbers (Study 1). Thus, while the processing and mental representations of numbers are highly automatic, their shape and strength depend on task demands. Both also seem to depend on language script direction and other cultural directional preferences. However, even with right-to-left language script, we did not observe a full reversal of the group-level SNARC effect (Study 3). Beyond numbers, SNARC-like effects extend to other ordinal stimuli. For letters, we found spatial mappings consistent with alphabetical order and, when made sufficiently salient, also with the order in newly memorized sequences (Study 4). Relatedly, while we replicated that the SNARC effect depends on relative magnitude or ordinality within the number set, it also seems to depend on absolute magnitude or ordinality (Study 5). Hence, the SNARC and SNARC-like effects are partly flexible, relying on working memory, and partly fixed, relying on long-term memory. Finally, we investigated the inter- and intraindividual variability of the SNARC effect. Only about half the sample showed a reliable SNARC effect in color judgment (Study 1), similar to prevalence rates reported for magnitude classification and parity judgment. Our novel longitudinal design (the “Ironman paradigm”) revealed that the temporal stability of the SNARC effect in parity judgment

was low and unsystematic (Study 6). Taken together, these studies simultaneously highlight the robustness and the variability of spatial mental representations of magnitude and ordinality. The SNARC and SNARC-like effects are shaped by task instructions, cultural background, and memory content, and are further characterized by inter- and intraindividual differences. Based on the findings of all studies, a computational contextualization model is introduced, building upon the assumption that the combination of simultaneously activated cognitive context schemata drives the SNARC and SNARC-like effects. Preliminary results obtained from initial data simulations and parameter optimization are promising, suggesting that the model may account for different shapes and strengths of the effects, as well as for their high variability.

Zusammenfassung (German)

Zahlen sind im Alltag fundamental, weshalb es essentiell ist, ihre kognitive Verarbeitung und mentalen Repräsentationen zu verstehen. Ein zentrales Phänomen im Bereich der numerischen Kognition ist der SNARC-Effekt (Spatial-Numerical Association of Response Codes), welcher schnellere Reaktionen auf kleine Zahlen mit der linken Hand und auf große Zahlen mit der rechten Hand beschreibt (Dehaene et al., 1993). Diese Dissertation befasst sich in sechs großen (überwiegend online durchgeführten) Studien mit verschiedenen Kontroversen rund um den SNARC-Effekt und ähnliche Effekte. Üblicherweise wird der SNARC-Effekt in beidhändigen Entscheidungsaufgaben mit Zahlenstimuli von 1 bis 9 (ohne 5) untersucht. Aufgabenrelevante Merkmale umfassen Zahlengröße oder -ordinalität (z.B. Größenurteile: kleiner vs. größer als 5), andere semantische Merkmale (z.B. Paritätsurteile: ungerade vs. gerade) oder nicht-semantische Merkmale (z.B. Farburteile: blau vs. gelb). Wir stellten einen kontinuierlichen Verlauf des SNARC-Effekts bei Paritätsurteilen fest (graduell wachsende Assoziation mit rechts relativ zu links mit zunehmender Zahlengröße), jedoch einen eher kategorialen Verlauf bei Größenurteilen (ähnliche linksseitige Assoziation für kleine und rechtsseitige Assoziation für große Zahlen; Studien 2 und 3). Trotz gemischter Vorbefunde konnten wir bei Farburteilen ebenfalls einen SNARC-Effekt messen, der allerdings schwächer ausgeprägt war und nur durch kleine Zahlen ausgelöst wurde (Studie 1). Demnach sind die Verarbeitung und mentale Repräsentationen von Zahlen zwar weitgehend automatisch, ihre Form und Stärke hängt jedoch von der Aufgabestellung ab. Letztere scheinen zudem durch die Richtung der Schriftsprache und andere kulturelle Richtungspräferenzen beeinflusst zu werden. Allerdings konnten wir selbst bei Sprachen mit Schriftrichtung von rechts nach links keine vollständige Umkehrung des SNARC-Effekts auf Gruppenebene erkennen (Studie 3). Abgesehen von Zahlen gelten SNARC-ähnliche Effekte auch für andere ordinale Stimuli. Bei Buchstaben fanden wir räumliche Verknüpfungen mit der alphabetischen Reihenfolge und, sofern diese ausreichend hervorgehoben wurden, auch mit der Reihenfolge in neu gelernten Sequenzen (Studie 4). Außerdem konnten wir replizieren, dass der SNARC-Effekt zwar von relativer Größe oder Ordinalität innerhalb der Zahlenmenge abhängt, jedoch auch von absoluter (Studie 5). Somit sind der SNARC-Effekt und ähnliche Effekte

teilweise flexibel, da sie auf dem Arbeitsgedächtnis beruhen, teilweise jedoch auch starr, da sie auf dem Langzeitgedächtnis beruhen. Zuletzt untersuchten wir die inter- und intraindividuelle Variabilität des SNARC-Effekts. Nur etwa die Hälfte der Stichprobe zeigte einen reliablen SNARC-Effekt bei Farburteilen (Studie 1), ähnlich der für Größen- und Paritätsurteile berichteten Prävalenz. Unser neues Longitudinaldesign (das „Ironman-Paradigma“) zeigte, dass die zeitliche Stabilität des SNARC-Effekts bei Paritätsurteilen gering und unsystematisch war (Studie 6). Insgesamt verdeutlichen diese Studien die gleichzeitige Robustheit und Varianz räumlicher mentaler Repräsentationen von Größe und Ordinalität. Der SNARC-Effekt und ähnliche Effekte werden durch Aufgabenstellungen, kulturellen Hintergrund und Gedächtnisinhalt geprägt und zeichnen sich zudem durch inter- und intraindividuelle Unterschiede aus. Auf der Basis aller Studienbefunde wird ein computergestütztes Kontextualisierungsmodell eingeführt. Dieses basiert auf der Annahme eines Zusammenspiels gleichzeitig aktivierter kognitiver Kontextschemata, welches den SNARC-Effekt und ähnliche Effekte hervorruft. Vorläufige Ergebnisse aus ersten Datensimulationen und Parameteroptimierungen sind vielversprechend und deuten darauf hin, dass das Modell unterschiedliche Formen und Stärken der Effekte sowie deren hohe Variabilität erklären kann.

Abbreviations

Abbreviations used in this dissertation (listed in the alphabetical order)

Abbreviation	Definition
AM	Absolute magnitude
ATOM	A Theory Of Magnitude
BAFT	Brain's Asymmetric Frequency Tuning
C0, C1, C2, C3, C4	Cognitive context schemata 0, 1, 2, 3, and 4
CDPQ	Cultural Directional Preferences Questionnaire
CJ	Color judgment
CVC	Consonant-vowel classification
dRT	Difference in RT between sides (RT right – RT left), usually averaged across all correct responses
LR	Left-to-right direction
LTM	Long-term memory
MARC effect	Linguistic Markedness of Response Codes effect
MC	Magnitude classification
MNL	Mental Number Line
OPE	Ordinal Position Effect
PJ	Parity judgment (odd vs. even)
RL	Right-to-left direction
RM	Relative magnitude
RMSE	Root mean squared error
RT	Response time
SD	Standard deviation
SNA	Spatial-Numerical Association
SNARC effect	Spatial-Numerical Association of Response Codes effect
WM	Working memory

List of publications

Study 1:

Roth, L., Caffier, J., Reips, U.-D., Cipora, K., Braun, L., & Nuerk, H.-C. (2025). True colors SNARC: Semantic number processing is highly automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 51*(12), 1875–1908. https://doi.org/10.1037/xlm0001431

Study 2:

Roth, L., Cipora, K., Overlander, A. T., Nuerk, H.-C., & Reips, U.-D. (under review). *Shape of SNARC: How task-dependent are Spatial-Numerical Associations? A highly powered online experiment* [manuscript for stage-2 Registered Report, after in-principle acceptance of stage-1 Registered Report by PCI-RR]. https://osf.io/jyrf3/files/8dac2

Study 3:

Bulut, M., Roth, L., Bahreini, N., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2025). One direction? Cultural aspects of the mental number line beyond reading direction. *Psychological Research, 89*, Article 37. https://doi.org/10.1007/s00426-024-02038-4

Study 4:

Roth, L.*, Huber, J. F.*, Kronenthaler, S., van Dijck, J., Cipora, K., Butz, M. V., & Nuerk, H.-C. (2025). Looks like SNARC spirit: Coexistence of short- and long-term associations between letters and space. *Quarterly Journal of Experimental Psychology, 78*(10), 2110–2132. https://doi.org/10.1177/17470218251324437

Study 5:

Roth, L., Caffier, J., Reips, U.-D., Nuerk, H.-C., Overlander, A. T., & Cipora, K. (2025). One and only SNARC? Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude. *Royal Society Open Science, 12*(1), Article 241585. https://doi.org/10.1098/rsos.241585

Study 6:

Roth, L.*, Jordan, V.*, Schwarz, S., Willmes, K., Nuerk, H.-C., van Dijck, J.-P., & Cipora, K. (2024). Don’t SNARC me now! Intraindividual variability of cognitive phenomena – Insights from the Ironman paradigm. *Cognition, 248*, Article 105781. https://doi.org/10.1016/j.cognition.2024.105781

* shared first authorship

1. Introduction

Is this the real [SNARC]? Is this just fantasy?
Caught in [response sides], no escape from reality
Open your eyes, look up to the [plot] and see
I'm just a poor [slope], I need no [number line]
Because I'm easy come, easy go, little high, little low
[Every] way the wind blows, [it does] really matter to me, to me

(Mercury, 1975; adapted[1])

Numbers are highly relevant in humans' everyday lives. For instance, humans need to process numbers when grocery shopping, counting objects, dealing with times and dates, or quantifying physical dimensions. We are confronted with numbers in various situations, like in school across all education levels, in a wide range of jobs, in many household activities at home, and even when going out or traveling. Understanding how the human mind processes and mentally represents numbers is therefore crucial.

Imagine you were to count birds sitting on a tree. Would you start counting on the left or on the right? Imagine you were to order boxes by the quantity of marbles they contain. Would you put the boxes with fewer marbles on the left or on the right? And if you wanted to decrease the volume of a speaker or adjust the air conditioning, would you go for pressing its left or right button if the buttons were unlabeled? Although in all these examples, the left and right options seem equally reasonable at first, according to findings in the domain of numerical cognition, you might exhibit a tendency to respond with the left in each of them. This is the case for many humans (at least if they are from Western cultures and if they are right-handed): They tend to count from left to right, associate small magnitude with the left and large magnitude with the right, and prefer to subtract intensity on the left while adding intensity on the right. As we shall see in this dissertation, such spatial biases in number processing arise in many situations.

[1] *Disclaimer:* The opening lines of this dissertation are adapted from the song "Bohemian Rhapsody" written by Freddie Mercury for Queen (original lyrics © Queen Music Ltd.) for scholarly purposes.

1.1. Spatial-Numerical Associations

Associating numbers with space is considered one of the most universal aspects of human number processing. Various related phenomena are grouped under the umbrella term Spatial-Numerical Associations (SNA). They have been referred to as the “key to hidden, deep properties of numerical representations and processes operating on them” (p.1, Cipora, Haman, et al., 2020) and are therefore extensively studied in the field of numerical cognition. A prominent SNA effect is the Spatial-Numerical Association of Response Codes (SNARC) effect, which refers to the phenomenon that responses to small or large numbers are faster with the left or right side, respectively. This phenomenon is interpreted as reflecting an association between number magnitude or ordinality and horizontal directional space. Unlike the examples given earlier that demonstrate a rather explicit SNA, the SNARC effect rather reflects an implicit SNA. The SNARC effect was first identified by Dehaene et al. in 1993 in a Western sample. Over the past three decades, it has been replicated in many other studies using a variety of task instructions, response setups, stimulus sets, modalities, contexts, and samples, as will be described in the following.

The most common setup to measure the SNARC effect involves presenting numerical stimuli on a computer screen and recording response times (RT) in a binary decision task. Participants are instructed to make fast and accurate judgments about a feature of the stimuli, responding with a left or right computer keyboard key with their index fingers. Early experiments were run in lab settings, but since the beginning of web-based experimentation and the recognition of its advantages (Reips, 2000), an increasing number of experiments have been conducted online (Krantz & Reips, 2017). Typically, stimuli are Arabic digits from 1 to 9 (excluding 5; Cipora, van Dijck, et al., 2019; Nuerk et al., 2004), with exemplary features to be judged in such tasks including number magnitude (i.e., small vs. large), number parity (i.e., odd vs. even), or non-semantic features of the digits like font color (e.g., red vs. green). However, SNA can also be demonstrated using other stimulus formats, such as dice patterns or written number words (Nuerk et al., 2004). Apart from varying task instructions and stimuli, response setups can also be modified: In addition to bimanual responses, the SNARC effect was also demonstrated in bipedal responses (Schwarz & Müller, 2006), unimanual responses (Riello & Rusconi, 2011), and tracking eye movements (Hesse & Bremmer, 2017; Huang, 2024; Schwarz & Keus, 2004). Moreover, beyond the visual modality, the SNARC effect has been shown in the auditory modality as well (Nuerk et

al., 2004). As concerns the study samples, the SNARC effect has not only been found in adults, but also in children (for overviews, see McCrink & Opfer, 2014; Patro et al., 2014), and not only in Western cultures and left-to-right (LR) script direction[2], but also in other cultures and right-to-left (RL) script direction (Dehaene et al., 1993; Shaki et al., 2009; Zebian, 2005). Even some animals seem to associate number magnitude with horizontal directional space (e.g., Gazes et al., 2017; Rugani et al., 2015).

Beyond numbers, different kinds of non-numerical magnitudes, such as physical size, conceptual size and weight associated with words' referents, economic value, luminance levels, and even visual speed have been shown to be mentally represented in a spatial way (for a review, see Macnamara et al., 2018). Evidence for this comes from behavioral, developmental, and neuroscientific studies (Cleland et al., 2020; Cohen Kadosh et al., 2005, 2008; Dalmaso & Vicovaro, 2019; Giuliani et al., 2024; Pinel et al., 2004; Ren et al., 2011; Richter & Wühr, 2022; Vicovaro et al., 2025; Wühr & Seegelke, 2018). Such SNARC-like effects are in line with "A Theory Of Magnitude" (ATOM; Bueti & Walsh, 2009; Walsh, 2003), which postulates similar processing mechanisms for space, time, and numbers, supporting the idea of a generalized magnitude system that encompasses even more dimensions. Consequently, the term "SNARC effect" has been extended to the "Spatial Quantity Association of Response Codes effect" (Löffler et al., 2022; Toomarian & Hubbard, 2018; Walsh, 2003). Besides magnitude, ordinality also seems to be mentally mapped onto space. Some researchers have even suggested that only ordinality, rather than magnitude, drives the SNARC effect for numbers (see Section 1.2.2. on the WM account).

To summarize, the SNARC effect appears to be a robust phenomenon when it comes to the mental representation of numbers, and such spatial representations also seem to extend to other stimuli featuring magnitudes or ordinality. The following section introduces theoretical accounts for the SNARC and SNARC-like effects.

1.2. Theoretical accounts of the SNARC effect

Since its discovery over three decades ago, many theoretical accounts have been proposed for the SNARC effect. In the following, I outline the most relevant ones for the current dissertation in detail, namely the Mental Number Line (MNL) account, the Working Memory (WM) account, as well as polarity and markedness accounts. I

[2] In the present dissertation, the term "script direction" is used, which is sometimes also referred to as "writing direction" or "reading direction."

thereafter briefly explain further theoretical accounts and the contribution of mathematical and computational models.

1.2.1. The Mental Number Line account

The idea of a MNL has already been documented by Galton as early as in 1880, and Restle (1970) later more explicitly suggested its alignment with a horizontal continuum from the left (small numbers) to the right (large numbers), at least in Western cultures. With the discovery of the SNARC effect in 1993, Dehaene et al. suggested that the effect might result from the LR orientation of the MNL. In the literature, the MNL is often claimed to be flexible, meaning that only the relevant interval in the mental representation is activated. Put differently, depending on the numbers appearing in a given context (see Study 5), one can zoom in and out dynamically (Pinhas et al., 2013).

The MNL still remains one of the most popular explanations for the SNARC effect. For instance, Tzelgov et al., (2015) referred to it when distinguishing between primitives and non-primitives of the numeration system. According to their distinction, primitives are numbers whose meaning is stored in long-term memory (LTM) and directly mapped onto the MNL (such as single-digit numbers in the Arabic system), whereas non-primitives are numbers whose meaning needs to be generated when required. Moreover, the LR or RL direction of the MNL is often discussed when investigating the reversal of the SNARC effect in some samples (e.g., with different script directions; see Study 3) and contexts (e.g., when new links between numbers and space are created before the task; see Fischer et al., 2010).

An interesting question in the field is whether the MNL is a trait (i.e., temporally stable and cross-situationally consistent) or a state (i.e., fluctuating across time points). On the one hand, the MNL and its direction are often treated like a relatively stable cognitive construct stored in LTM. For instance, this is the case in investigations of correlations between the SNARC effect (assuming that it reflects a relatively stable underlying MNL) and other relatively stable variables, such as math skills (Cipora, He, et al., 2020). On the other hand, the SNARC effect depends on situational factors (Cipora et al., 2018), such as perceptual influences (e.g., horizontal location of numbers in texts read prior to the task; Fischer et al., 2010), representational influences (e.g., imagining numbers on a clock face versus on a ruler; Bächtold et al., 1998; Mingolo et al., 2023), or action influences (e.g., crossing hands; Dehaene et al., 1993). This suggests that the MNL is not always equally strong and can even be reversed

within individuals, rather reflecting a state (see Study 6). Note that modern personality models suggest that the manifestation of personality constructs depends on both the underlying personality traits and the current situational state (e.g., Endler & Kocovski, 2001; Steyer et al., 1999; Wilson et al., 2017). Thus, the MNL might have both trait-like and state-like elements.

To conclude, the MNL account postulates that numbers are mentally represented on a horizontal continuum. Its scope is limited to numbers and no generalization to other kinds of magnitude or ordinality is intended. While the MNL seems to entail both trait-like and state-like characteristics, it is most often conceived as an LTM concept.

1.2.2. The Working Memory account

In contrast to the MNL account, which suggests a mental representation of numbers in LTM, the WM account proposes that the mental representation of numbers is constructed in WM during task execution (Fias & van Dijck, 2016; van Dijck et al., 2014; van Dijck & Fias, 2011). According to this influential account, the SNARC effect is an instance of the Ordinal Position Effect (OPE; Ginsburg et al., 2014; which is also sometimes referred to as Spatial-Positional Association of Response Codes effect; Guida & Lavielle-Guida, 2014). The OPE describes a spatial mapping of items based on their ordinal position in a sequence (i.e., leftward/rightward shift in the mental representation of early/late items, respectively). As recently explained by Abrahamse and van Dijck (2023), such a sequence can be thought of as a “ranking-space,” where the ordinal position of each item is its rank relative to other items or reference points (e.g., the starting, mid, or endpoint of the sequence) in the ranking-space. On the one hand, evidence for the OPE comes from experiments in which participants must memorize arbitrary ordinal sequences before performing the experimental task. A spatial mapping of the ordinal positions stored in WM has been demonstrated using various stimuli, including numbers and names of fruits and vegetables (van Dijck & Fias, 2011), object names (Previtali et al., 2010), or newly created symbols (Van Opstal et al., 2009). On the other hand, although more commonly used to refer to the spatialization of WM ordinality, the term OPE can also be employed to refer to the spatialization of LTM ordinality. Evidence for the OPE based on LTM ordinality has been found for items that are part of canonical sequences, such as letters in the alphabetical order, days in the order of the week, or months in the order of the year (Gevers et al., 2003, 2004). When investigating spatial mappings of ordinality,

disentangling whether ordinality is retrieved from WM or LTM (see Study 4) is crucial. Importantly, the WM account considers both cases. Regarding the exact cognitive mechanism, Abrahamse et al. (2016) proposed a spatial template for ordinal information stored in LTM, which is retrieved and transferred to WM during task execution. Items are filled into this template within WM, but as explained above, the ordinality itself may originate from WM or LTM. In contrast, Guida and Campitelli (2019), in their Expertise account, argue that the association between ordinality and space is stored in LTM if the ordinality comes from LTM, and in WM if the ordinality comes from WM. These views are still not unified today.

As opposed to other stimuli exemplified above that entail ordinality (either newly learned as in names of fruits, or canonical, as in letters), numbers additionally entail magnitude, which is also referred to as cardinality. Although related, ordinality and cardinality are conceptually distinct (Lyons et al., 2016), as evidenced, for instance, by their acquisition at different time points (Spaepen et al., 2018). Ordinality refers to a position within a sequence conveyed by a number, whereas cardinality refers to the concept of quantity (Wasner et al., 2015). While acquisition of the "Cardinality Principle" (i.e., knowing that a set size equals to the last number word reached when counting items in the set) is considered a milestone in numerical cognitive development (Wynn, 1990), ordinality might play a less important role in math education. As explained above, ordinal sequences can be stored in WM or LTM, whereas the cardinal meaning of numbers is stored in LTM. The WM account's assumption that ordinality rather than cardinality drives the SNARC and SNARC-like effects is highly debated in the field. Casasanto and Pitt (2019; see also Pitt & Casasanto, 2022) support this view. Pitt outlined several reasons for this claim in a workshop talk ("The role of order and magnitude in Spatial-Numerical Associations," Tübingen, 2024). First, all stimuli that elicit SNARC-like effects entail ordinality (i.e., numerical and non-numerical stimuli), whereas magnitude seems optional (e.g., letters and months entail only ordinality)[3]. Secondly, a SNARC-like effect based on newly learned ordinality arises even when magnitude is task-relevant (Ginsburg et al., 2014; van Dijck & Fias, 2011), making magnitude seem weaker in its mental representation. Thirdly, the SNARC effect has been claimed to reflect the mental representation of the relative magnitude (RM) of

[3] In this dissertation, the term "SNARC-like effects" will be used for all effects referring to the spatial mapping of non-numerical magnitude or ordinality. This broad category includes both Spatial Quantity Association of Response Codes effects, referring to spatial mappings that are based on non-numerical magnitudes, and OPE, referring to spatial mappings that are based on ordinality.

numbers relative to other numbers in a set, rather than their absolute magnitude (AM; Dehaene et al., 1993; but see Study 5). RM within a set can, in fact, be interpreted as ordinality, as suggested by the idea of a ranking-space (Abrahamse & van Dijck, 2023): Smaller numbers usually appear earlier in a set than larger numbers.

Conversely, Prpic et al. (2016, 2021) claimed that the SNARC and SNARC-like effects can be based on either magnitude or ordinality, depending on task instructions (for another mixed view, see Schroeder et al., 2017b). In a thoughtful design, they used musical note values as non-numerical stimuli, in which magnitude (e.g., quarter, half, whole note) and ordinality (e.g., whole, half, quarter note) are opposed. They found that the SNARC effect is driven by ordinality in *direct* tasks, in which instructions are magnitude- or order-relevant, such as note-value comparison. Conversely, the SNARC effect appeared to depend on cardinality in *indirect* tasks, in which instructions are magnitude- or order-irrelevant, such as note detection in their study. In the case of number stimuli, magnitude classification (MC) is a common direct task, whereas parity judgment (PJ) is a typical indirect task. Note, however, that other authors have reported an OPE in both direct (order-relevant) and indirect (order-irrelevant) tasks (Gevers et al., 2003, 2004; Previtali et al., 2010).

Finally, the notion of a generalized magnitude system suggests that the SNARC and SNARC-like effects are based on magnitude rather than on ordinality. Namely, ATOM (Walsh, 2003), at least in its original form, suggests a relationship between prothetic continua (i.e., domains in which humans experience *quantitative* variations, e.g., physical size, color intensity; “how much” according to Stevens, 1957) and their association with space. However, ATOM does not make any statements about metathetic continua (i.e., domains in which humans experience only *qualitative* variations, e.g., letters in alphabetical order; “what kind” according to Stevens, 1957).

To summarize, SNARC-like effects for non-numerical magnitudes and ordinality have been observed, next to the SNARC effect referring only to the mental representation of numbers. However, it remains unclear under which conditions the SNARC and SNARC-like effects are based on magnitude or ordinality. The revised version of the WM account (Abrahamse et al., 2016) claims that they are based on ordinality, stemming either from WM or from LTM, but disregards the role of magnitude. While these distinctions are crucial for the WM account, they are not the central aspect of the polarity and markedness accounts introduced below.

1.2.3. Polarity and markedness accounts

A different approach to explaining the SNARC effect is taken by the following three accounts: the Polarity Correspondence principle (Proctor & Cho, 2006), the Linguistic Markedness account applied to numerical cognition (Nuerk et al., 2004), and the newly developed Psychological Markedness account (Nuerk & Schroeder, 2024). The first two can be summarized as verbal-spatial accounts, in comparison to rather than visuo-spatial accounts such as the MNL account (for a differentiation between verbal-spatial and visuo-spatial accounts, see Gevers et al., 2010). According to all three accounts, the SNARC effect is a compatibility effect between the polarity or markedness status of stimuli and response sides. Specifically, stimuli and response sides are assumed to entail either positive polarity (i.e., positive affective valence, pleasant; Proctor & Cho, 2006) or negative polarity (i.e., negative affective valence, unpleasant). At the same time, they can also be categorized as "unmarked" (i.e., simple, common, or acquired early; Haspelmath, 2006; Hume, 2004) or "marked" (i.e., complex, uncommon, or acquired late). Importantly, polarity and linguistic markedness statuses often coincide: Verbal stimulus labels like "large" and "even" are claimed to be coded with a positive polarity and categorized as unmarked, just like the verbal response label "right." In contrast, "small" and "odd" are assumed to be coded with a negative polarity and described as marked, just like "left". When applying the linguistic markedness concept to numerical cognition, Nuerk et al. (2004) also introduced the MARC effect (Linguistic Markedness of Response Codes effect; i.e., the association between odd/even number parity status and left/right response side, respectively). This seminal paper explains the role of formal markedness (e.g., in German, the literal translation of the marked adjective "odd" is "uneven," which is the negation of the unmarked ground form "even"[4]) and distributive markedness (e.g., the unmarked word "right" is more frequently used than the marked word "left"). Such linguistic asymmetries can lead to cognitive asymmetries like compatibility effects or general speed differences when processing opposite numerical features (e.g., the Odd effect, reflecting faster responses to even than odd numbers; Hines, 1990, 2010), producing typical numerical-cognition effects. Hence, the polarity and linguistic markedness accounts predict, among other effects, a small-left and large-right compatibility effect (i.e., the SNARC effect), and an odd-left and even-right compatibility effect (i.e., the

[4] Paradoxically, the term "marked" is the ground form and therefore unmarked, whereas the term "unmarked" is its negation and therefore marked.

MARC effect), reflecting the mental spatialization of magnitude and parity. Accordingly, pure verbal-spatial coding may account for such effects and visuo-spatial coding is not required for an explanation (Gevers et al., 2010). The fact that the MARC effect is found more consistently in number words than in Arabic digits (Nuerk et al., 2004; Roettger & Domahs, 2015) lends additional support to linguistic explanations. Note that number magnitude is continuous, while parity is dichotomous. Hence, if two opposing categorical verbal labels are responsible for the effects, a dichotomization of magnitude is required to elicit a SNARC effect, whereas no dichotomization of parity is required to elicit a MARC effect. Whether verbal labels induce or only reinforce SNA remains under debate (Toomarian & Hubbard, 2018).

The Psychological Markedness account (Nuerk & Schroeder, 2024) was originally based on the Linguistic Markedness account (see also Schroeder et al., 2017c). Therein, the authors define psychological markedness as relying on mental representations, being graded and embodied, differing between individuals, and depending on the current context. With regards to the SNARC and MARC effects, the Psychological Markedness account suggests that they are driven by mental representations of magnitude and parity, rather than mathematical or linguistic features of magnitude and parity. This implies that the psychological markedness may be graded, although it is dichotomous for features like “odd” and “even” from mathematical and linguistic perspectives. For instance, prime numbers seem to be perceived as “more odd” than other odd numbers, and powers of 2 as “more even” than other even numbers (Heubner et al., 2018), suggesting that the mental representation of parity is a continuum rather than a dichotomy. Moreover, psychological markedness of the same features or stimuli may differ between individuals. For instance, while “right” is unmarked and “left” is marked from a linguistic perspective, this might not be true for each individual from a psychological or embodied perspective: For left-handers, who execute actions with the left hand as a default side, “left” might be unmarked and “right” might be marked. This is in line with findings like that the MARC effect is more reversed the more left-handed participants are (Cipora, Soltanlou, et al., 2019; Huber et al., 2015). Psychological markedness is also thought to depend on the current context, leading to intraindividual differences based on factors such as notational modulations of stimuli or the surrounding language context. For example, psychological markedness of parity may be more pronounced in verbal stimuli compared to digits (Nuerk et al., 2004), and shift depending on the language in which a bilingual individual

has just read a text (Shaki & Fischer, 2008). Moreover, beyond compatibility effects between stimuli and response sides (e.g., SNARC effect, MARC effect), the Psychological Markedness account also explains compatibility effects between different stimuli. An example of such an effect is the Parity-Numerical Association of Representational Codes (Krajcsi et al., 2018) effect, which describes an association between small/large number magnitude and odd/even parity statuses.

To conclude, the Polarity Correspondence principle and the Linguistic Markedness account differentiate between two broadly used verbal labels for stimuli and response sides. When applying these theories to number processing, cognitive phenomena such as the SNARC and MARC effects may be caused by compatibility of the polarity or markedness between verbal labels of stimuli and response sides. The Psychological Markedness account, however, goes much beyond verbal label compatibility. It explains cognitive phenomena with mental representations and allows for differences between individuals and contexts.

1.2.4. Other theoretical accounts

The accounts introduced above may be unable to explain all SNARC-related findings, at least in their most prominent forms. For example, the MNL account only justifies the high degree of flexibility of the SNARC effect when assuming that zooming in and out is possible, and it only holds for numbers but not for non-numerical stimuli. The WM account fails to explain long-term associations, and it does not predict well when the spatial template proposed for short-term associations is filled in with content from WM or LTM. The Polarity Correspondence principle and the Linguistic Markedness accounts only consider dichotomous features, and they do not account for the flexibility and variability of the SNARC effect, both within and between individuals. Because of such limitations, a hybrid account has been proposed (van Dijck et al., 2015), integrating views from multiple accounts (i.e., the MNL account, the WM account, and the Polarity Correspondence principle). The hybrid account suggests coexisting mental representations, which are activated depending on the cognitive processing mechanisms that the executed task requires. Later on, a multiple-coding framework has been proposed, including even more different sources of the SNARC effect (Schroeder et al., 2017c). While some sources are verbal (i.e., linguistic or psychological markedness, and verbal-spatial instructions), others are non-verbal (i.e., visuospatial simulation, and ordinality of sequences stored in WM). These verbal and

non-verbal codes are activated depending on factors like the task, stimuli, and culture, but statements about the interplay of different codes remain vague.

Apart from the accounts described above, further theoretical accounts have been proposed to explain the SNARC effect, such as the dual-route model, and the Brain's Asymmetric Frequency Tuning (BAFT) hypothesis. These two accounts will be briefly described below. Finally, I will outline the importance and usefulness of mathematical and computational models for cognitive effects such as the SNARC effect.

The dual-route model (Gevers, Ratinckx, et al., 2006; Gevers, Verguts, et al., 2006), which has been implemented as a computational model, suggests that the SNARC effect arises as a compatibility effect between the outcomes of two parallel processing routes. Responses are fast when the unconditional route, which activates an automatic stimulus-to-response mapping, converges with the conditional route, which activates an intentional stimulus-to-response mapping according to the task instructions. For instance, the number 1 activates a left-hand response according to the unconditional route because of learned connections between verbal magnitude labels (i.e., "small" and "large") and spatial representation labels (i.e., "left" vs. "right"). Such verbal-spatial connections (Gevers et al., 2010) are in line with polarity or markedness accounts explained earlier (Nuerk et al., 2004; Nuerk & Schroeder, 2024; Proctor & Cho, 2006). The activation of a left-hand response for number 1 is compatible with an odd-left and even-right instruction in PJ. In contrast, an even-left and odd-right instruction in PJ would make trials with the stimulus number 1 be incompatible. Here, the two routes lead to different stimulus-to-response mappings, thus producing slower responses.

Furthermore, the BAFT hypothesis suggests that a lateralized nature of frequency processing leads to asymmetries between the hemispheres (Felisatti, Aagten-Murphy, et al., 2020), which in turn causes SNA (Felisatti, Laubrock, et al., 2020a). Specifically, the right hemisphere favorably processes low spatial frequencies, which correspond to small quantities, while the left hemisphere favorably processes high spatial frequencies, which correspond to large quantities. The BAFT hypothesis therefore predicts a left-side behavioral bias for small quantities and a right-side behavioral bias for large quantities. This prediction is in line with the SNARC effect and does not require additional cognitive mechanisms (Felisatti, Laubrock, et al., 2020b).

Finally, a different approach for uncovering cognitive mechanisms that drive the SNARC effect is mathematical or computational modeling. Such models can help bridge the gap between observations and underlying factors, and they are therefore considered indispensable to cognitive science (Farrell & Lewandowsky, 2018, p. 3; Ulrich, 2009). Verbally expressed theories often rely on intuition, sometimes include inconsistencies or logical contradictions, and may be weaker in comparison to mathematical and computational models (Fum et al., 2007; Ulrich, 2009). An influential and often-cited computational model for the SNARC effect was introduced by Gevers, Verguts, et al. (2006). It is based on a model of exact small-number representation (Verguts et al., 2005) and on the dual-route model (Gevers, Ratinckx, et al., 2006). The computational model consists of three layers. The bottom layer represents the MNL, consisting of a number field with one node per number and a standard field coding the reference number (if existent, as in MC) or the mean number magnitude (if no reference number exists, as in PJ) depending on the task instructions. The middle layer receives input from the bottom layer. It always includes a magnitude field with a small and large node but can also include additional fields depending on the task instructions, such as a parity field with an odd and even node in PJ. The top layer consists of a response field with a left and right node, leading to a left or right response according to the first node to hit the threshold. A dual-route architecture determines how fast and how accurately the threshold is reached (De Jong et al., 1994), depending on the convergence of a fast unconditional route that is automatically activated and a slow conditional route that depends on task instructions. Model predictions, such as a continuous SNARC effect in PJ but a categorical SNARC effect in MC, were confirmed by two experiments.

Crucially, the computational model by Gevers, Verguts, et al. (2006) reveals several limitations. First, it does not include a middle-layer field for more precise automatic processing of small numbers compared to large numbers, which we consider to be necessary given the stronger spatial mental representations of small numbers (for significant differences, see Studies 1 and 5; for descriptive indications, see Studies 2 and 3). Moreover, this computational model does not account for interindividual differences in the SNARC effect. It has not yet been generalized to ordinal stimuli other than numbers (e.g., letters) either and does therefore not make any predictions for SNARC-like effects. We therefore created a new computational model with a simple yet flexible architecture (see Section 3.3.). It grounds on the idea

of contextualization (Butz et al., 2025), assuming that the simultaneous activation of several cognitive context schemata is responsible for the SNARC and SNARC-like effects. These schemata are defined by spatial biases in processing, which reflect behavioral patterns we observed in our studies, as well as different aspects of the theoretical accounts described above.

Having explained the most dominant theoretical accounts for the SNARC effect, I will next introduce different influences on the SNARC effect. This will then lead to the hypotheses and paradigms of the six studies conducted in this dissertation.

1.3. Influences on the SNARC effect

In the following, different tasks requiring automatic or intentional processing of numbers, different cultures and languages, different stimulus sets used in experiments, and inter- and intraindividual variability will be introduced as potential influences on the SNARC effect.

1.3.1. The SNARC effect in automatic and intentional processing tasks

In the field of cognition, researchers differentiate between automatic and intentional processing. In contrast to intentional processing, automatic processing is uncontrolled, unconscious, efficient, and fast (Moors, 2016). While some early influential theories took an all-or-none approach to automaticity (e.g., Bargh, 1994), others supported a more gradual view (e.g., Treisman et al., 1992; for a review, see Moors & De Houwer, 2006). Nowadays, the gradual view is more widely accepted, conceptualizing automaticity and its features as a matter of degree (Moors, 2016). This view will therefore be adopted in the present dissertation.

What makes the SNARC effect interesting for numerical cognition research is that it reflects a certain degree of automaticity when it comes to processing number and building mental representations for them (for automatic versus intentional processing of numbers, see Tzelgov et al., 1992). First, automaticity can be assumed because the SNARC arises in experimental tasks like MC that require participants to process number magnitude, but not to relate it to space (Dehaene et al., 1993; Wood et al., 2008). Note that the LR dimension is primed in experiments using a horizontal response-key setup (Shaki & Fischer, 2018), although this seems not to suffice for producing a SNARC effect (Pinto et al., 2019, 2021). Second, this even more obvious with respect to the typically observed SNARC effect in tasks where number magnitude is entirely task-irrelevant, because only other semantic features of numbers are judged,

such as in PJ (Dehaene et al., 1993). Odd and even numbers have the same average magnitude (i.e., 5) in the stimulus set that is typically used for PJ (i.e., 1 to 9, excluding 5), so processing their magnitude and associating it with space do not provide any obvious advantage in performing PJ. Third and most convincingly, the SNARC effect has even been demonstrated in at least some tasks requiring only the processing of non-semantic (i.e., superficial visual) features. Such tasks were orientation judgment tasks involving triangles superimposed on digits (pointing upwards vs. downwards; Fias et al., 2001, Exp. 1), lines superimposed on digits (horizontal vs. vertical; Fias et al., 2001, Exp. 4), or digits themselves (upright vs. tilted by 10 degrees; Lammertyn et al., 2002; but see Wang et al., 2024, for a non-significant result when judging tilted left vs. right by 45 degrees). As an interim summary, automatic number processing and automatic building of mental number representations can be observed in tasks that do not require participants to relate numbers to space or even process them semantically in the first place.

Interestingly, in contrast to more robust evidence for the SNARC effect in orientation judgment, evidence for the SNARC effect in color judgment (CJ) is scarce. In most studies, it has neither been detected in *nominal* CJ, where participants were tasked with differentiating color hues (red vs. green: Didino et al., 2019; Fias et al., 2001, Exp. 2; Keus & Schwarz, 2005, Exp. 3; blue vs. green: Cleland & Bull, 2019, Exp. 1 and 5; Mitchell et al., 2012, Exp. 1, and Cleland & Bull, 2025, with non-symbolic dot arrays; Schroeder et al., 2017, with digits 1, 2, 4, 5 only; blue vs. yellow: Schroeder et al., 2017b, with digits 1, 2, 4, 5 only), nor in *intensity* CJ (light cyan vs. dark cyan; Fias et al., 2001, Exp. 3; Lammertyn et al., 2002). To my knowledge, the SNARC effect was only found in nominal CJ in 5-year-old children (Hoffmann et al., 2013), though only after completing MC, which might have primed magnitude processing. Moreover, it was found in a CJ task in a recent study (Lengyel & Krajcsi, 2024); however, the numbers were first presented in black, enabling the participants to process magnitude before they could perform the task, as the color onset occurred only after 400 ms. Only a SNARC-like effect for physical size has been found in nominal CJ (Wühr & Seegelke, 2018). To summarize, the SNARC effect seems to be present in orientation judgment, while studies yielded mixed evidence in CJ, with most of them not reporting any SNARC effect.

Early on, Fias et al. (2001) formulated the “neural overlap hypothesis” to explain these findings, arguing that the SNARC effect only arises if there is sufficient neural

overlap of structures dedicated to processing relevant (e.g., orientation in orientation judgment, color in CJ) and irrelevant (i.e., number magnitude) information. As the parietal cortex is highly involved in processing number magnitude and orientation (dorsal pathway), but not in processing color (ventral pathway), the SNARC effect should, according to their hypothesis, arise only in orientation judgment but not in CJ. Nevertheless, subsequent work in the field has offered arguments against this hypothesis (Keus, 2006; Toomarian & Hubbard, 2018). Namely, neural overlap between number and color processing has been demonstrated (Claeys et al., 2004), and the SNARC effect might arise only during the response selection stage rather than during the stimulus processing stage (Keus et al., 2005; Keus & Schwarz, 2005).

Other potential reasons for the null results regarding the SNARC effect in CJ in the aforementioned articles are methodological in nature. Most importantly, only frequentist analyses were run, making the null results hardly interpretable. In addition, both sample sizes and numbers of trials were relatively small in most experiments. In Study 1, we therefore aimed to replicate two CJ experiments in an online format. This approach allowed us to test large sample sizes and thereby achieve high statistical power even for the detection of small effects (Reips, 2000, 2002). We also implemented a sufficient number of trials per experiment, carefully controlled for potentially confounding variables, and calculated Bayes Factors alongside frequentist analyses to quantify evidence of absence. Because CJ, like orientation judgment, requires neither semantic processing of number magnitude nor the activation of space, observing a SNARC effect in CJ would suggest that semantic processing of number magnitude is even more automatized than previously assumed.

In contrast to the aforementioned tasks requiring judgments on non-semantic features, evidence for the SNARC effect is highly robust in tasks requiring judgments on semantic features. The two most widely used tasks requiring judgments on semantic features are MC (i.e., judging whether numbers between 1 and 9, excluding 5, are smaller or larger than 5) and PJ (i.e., judging whether numbers are odd or even). Although seemingly similar at first sight, MC and PJ differ in several aspects. Most importantly, because number magnitude is task-relevant to MC but not to PJ, its processing is intentional in MC but automatic in PJ (as classified by Tzelgov et al., 2015). One might therefore expect a stronger SNARC effect in MC than in PJ. Moreover, the processing of number magnitude and its association with space seem to differ between MC and PJ, which becomes apparent when comparing the shapes of

the SNARC effect. Specifically, in MC, the left-side association seems equally strong for all numbers from 1 to 4, and the right-side association seems equally strong for all numbers from 6 to 9. By comparison, in PJ, the spatial mapping of extreme magnitudes (e.g., 1 and 9) seems stronger than that for magnitudes in the middle of the stimulus set (e.g., 4 and 6). In other words, the SNARC effect seems to be categorical in MC and continuous in PJ (Wood et al., 2008), a phenomenon we aimed to replicate in Study 2.

Apart from the SNARC effect, the MARC effect also seems to differ between MC and PJ. Importantly, number parity is task-relevant to PJ but not to MC. The fact that the MARC effect was only detected in PJ but not in MC in previous studies (Cipora, 2014; Deng et al., 2018) leads to the interpretation that parity processing is less automatic than magnitude processing. Replicating this pattern was another aim of Study 2, which used a much larger online sample than previous studies. Furthermore, because the MARC effect has not been reported in previous articles about CJ tasks, we tested it in Study 1. Because it is typically absent even in MC, which requires semantic number processing, we did not expect to find the MARC effect in CJ, which requires only superficial visual processing of the presented numbers.

To summarize, many studies have demonstrated automatic number magnitude processing and its automatic association with space across a variety of tasks. Interestingly, the SNARC effect is even detectable in some tasks where numbers do not need to be semantically processed at all. Since evidence regarding the presence of the SNARC effect in CJ tasks is inconsistent, but null findings mostly stem from small samples, we conducted both nominal and intensity CJ with large samples in Study 1, hypothesizing a SNARC effect in both tasks. The SNARC effect is found in most studies employing judgment tasks of semantic number features. However, the field still lacks a systematic investigation of its shape, a determination of the most suitable analysis method, a comparison of its strength, and a test of its correlation between MC and PJ. To address this gap, we therefore let a large sample complete the two most common semantic SNARC tasks online in a within-subjects design in Study 2.

1.3.2. The SNARC effect in different cultures and languages

As pointed out by Toomarian and Hubbard (2018), only humans use formal reading and writing systems, differing between cultures, and it is thus not farfetched that these systems shape human mental representations, unlike in other species.

Importantly, SNA have mostly been tested in Western samples. Their languages follow a LR script direction, and their MNL seems to be LR as well. The question arising from this congruency is whether script direction predicts the orientation of the MNL, as suggested in the seminal study by Dehaene et al. (1993). In their Experiment 7, the authors found no SNARC effect in Iranian participants living in France. This was attributed to the simultaneous use of the RL script direction of the Farsi (i.e., Persian) language and the LR script direction of the French language. An additional analysis revealed that the regular LR SNARC effect was stronger the longer participants had been in France. In a subsequent study with Lebanese participants, Zebian (2005) investigated the SNARC effect in Arabic, a language with a RL script direction. She found a reverse RL SNARC effect for Arabic monoliterates, a weaker reverse RL SNARC effect for Arabic-English or Arabic-French biliterates, and no significant effect for Arabic illiterates. In a cross-cultural study, Shaki et al. (2009) investigated the SNARC effect systematically by recruiting samples from three countries: Canadians, whose script direction is LR for words as well as for Arabic numbers; Israelis, whose script direction is RL for words but LR for Arabic numbers; and Palestinians, whose script direction is RL for words as well as for Eastern Arabic numbers. In line with their respective script directions, Canadians showed a regular LR SNARC effect, no significant effect was found for Israelis, and Palestinians showed a reverse RL SNARC effect. Importantly, however, these early studies have several methodological limitations. First, sample sizes were small, and the statistical power was therefore insufficient for the detection of small effects. Second, the number of repetitions for each stimulus in each response-to-key assignment was too low, resulting in compromised measurement precision (see Cipora & Wood, 2017, for recommendations).

Interestingly, the reversal of the SNARC effect in participants with a RL language has not been replicated in newer studies. Instead, a regular LR SNARC effect was found in Israeli Hebrew-speaking children (Feldman et al., 2019), in Israeli Hebrew-speaking adults (although only when reducing the MARC effect; Zohar-Shai et al., 2017), in Jordanian Arabic-speaking adults (although weakened compared to US-American English-speaking adults; Lopiccolo & Chang, 2021), in Iranian Farsi-speaking adults (Hochman et al., 2024), and in Farsi-speaking adults (Cipora, Soltanlou, et al., 2019, Supplementary Materials). In line with this, other SNA like the association of subtraction with the left and addition with the right do not seem to differ between French- and Arabic-speaking adults either (Masson et al., 2020).

Apart from differences between individuals, script direction also seems to influence the SNARC effects within individuals. For example, Taiwanese Chinese-speaking adults show a LR SNARC effect in PJ with Arabic numbers (always written LR) using a horizontal response-key alignment, but not with Chinese number words (mostly written top-to-bottom). At the same time, they reveal a vertical top-to-bottom SNARC effect with Chinese number words, but not with Arabic numbers (Hung et al., 2008). In a similar vein, contextual influences of script direction have been found in two studies with Russian-Hebrew bilinguals). In one study, the SNARC effect was regular after reading a paragraph in Russian (LR language), whereas no effect was found after reading a paragraph in Hebrew (RL language; Shaki & Fischer, 2008). In another study, even a single number word was sufficient to manipulate the mental spatialization of number magnitude: A regular SNARC effect in Arabic digits was only found after Russian number words, while no effect was found after Hebrew number words (Fischer et al., 2009).

The influence of script direction on spatial associations has also been investigated with non-numerical ordinal information stored in WM. In line with the WM account (van Dijck et al., 2014; van Dijck & Fias, 2011), multiple studies have shown that the direction in which ordinal information is stored in WM corresponds to the script direction (e.g., LR in French speakers, Guida et al., 2015; RL in Arabic speakers, Guida et al., 2018; RL in Farsi speakers, Rasoulzadeh et al., 2024). Notably, this direction in WM can apparently be intentionally reversed if requested (Guida et al., 2020). A similar pattern is observed for spatial associations of face age (Dalmaso et al., 2023): In a cross-cultural study, Italians responded faster to younger faces on the left and to older faces on the right, whereas the pattern was reversed in Iranians (Erfanikia, 2024).

Many of the results mentioned above show an influence of script direction on SNA and hence suggest cultural modulations of SNA. However, it is important to note that there are more cultural directional preferences than language script direction alone. For instance, adults and even preschoolers preferably count horizontally aligned objects in their script direction (e.g., Opfer & Furlong, 2011; Shaki et al., 2012). Similarly, effects have been found for horizontal ordering of temporal stimuli or movements such as drawing horizontal lines (for an overview, see Nuerk et al., 2015, and Patro et al., 2016). According to Patro, Nuerk, et al. (2016), cultural directional preferences influence SNA already in the preliterate age, when children make directional experiences, for example by observing directional movements performed

by their parents. Naturally, the direction of such cultural preferences usually aligns with the script direction (Faghihi & Vaid, 2023). However, this alignment is not always the case. For instance, in Turkey, the RL script was replaced with a LR script in 1928 (Yılmaz, 2011; note that according to that source, as few as 8% of the Turkish population were literate in 1927, though), but former cultural RL habits may still persist. Additionally, in Islam, a prominent religion in Turkey, adherents are encouraged to begin actions with the right rather than the left. A Turkish population therefore provides a unique opportunity to disentangle script direction from other cultural directional preferences. If both script direction and other cultural directional preferences influence SNA, the SNARC effect in a Turkish sample should be weaker than in a sample where both directions are LR, but stronger than in a sample where both directions are RL. In line with this prediction, most studies with Turkish samples do not find any SNARC effect in PJ (Bulut et al., 2023; Kaya et al., 2025) and MC (Bulut et al., 2025, 2025; Dural et al., 2025).

Crucially, the spatial mental representation of number magnitude is not necessarily shaped solely by environmental factors such as cultural directional preferences, although even archaeological and historical findings suggest this (Núñez, 2011). Instead, some researchers argue that SNA may be innate. One line of evidence for this claim comes from studies with preverbal humans (i.e., infants prior to speech development and reading or writing skills). Even 0- to 3-day old neonates seem to exhibit a preference for LR over RL sequences of increasing duration, numerical magnitude, and auditory magnitude (de Hevia et al., 2014, 2017). They associate numbers smaller or larger than a reference with the left or right, respectively (di Giorgio et al., 2019). Another line of evidence comes from similar studies with non-human species. For example, LR counting behavior has been demonstrated for newborn chicks, adult nutcrackers, rhesus monkeys, and honey bees (Drucker & Brannon, 2014; Giurfa et al., 2022; Rugani et al., 2010). Moreover, newborn chicks, gorillas, and orangutans seem to relate smaller magnitudes to the left and larger ones to the right (Gazes et al., 2017; Rugani et al., 2015). That said, no SNARC effect – or only limited evidence for it – was detected in several other studies with rhesus monkeys, capuchin monkeys, western lowland gorillas, American black bears, and cleaner fish (Beran et al., 2019; Johnson-Ulrich & Vonk, 2018; Silva et al., 2025; Triki & Bshary, 2018). In sum, while there is evidence that the SNARC effect is shaped by culture, studies with preverbal humans and non-human animals also suggest that it may be innate (for a

review, see Rugani & de Hevia, 2017). Theories incorporating both views, such as spatial biases being innate at first and shifting to being enculturated during childhood, have also been suggested (McCrink & de Hevia, 2018).

To summarize, there is an ongoing debate about whether the SNARC and SNARC-like effects are innate or shaped by culture, and evidence exists to support both views. Regarding cultural influences, previous research has shown that the SNARC effect aligns with a LR MNL in participants most proficient in LR languages, whereas findings are inconclusive for participants being most proficient in RL languages. In addition to language, other cultural directional habits might play a role in how mental representations of numbers are spatialized. In Study 3, we therefore ran a cross-cultural investigation of the SNARC effect with samples coming from three countries, using the two most common SNARC tasks as in Study 2. Namely, we investigated the SNARC effect in both MC and PJ in German (LR language and predominantly LR habits), Turkish (LR language and potentially more RL habits), and Iranian (RL language and predominantly RL habits) participants.

1.3.3. The SNARC effect in different stimuli

Apart from recruiting different samples, as described in the previous section, experimenters can also investigate the SNARC and SNARC-like effects using different kinds of stimuli. The most common set consists of numbers from 1 to 9 (excluding 5), where ordinal position and magnitude almost perfectly correspond to one another (i.e., the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th numbers are 1, 2, 3, 4, 6, 7, 8, 9), making them difficult to disentangle (see Koch et al., 2023a, for an approach). In classic SNARC investigations, no arbitrary sequence must be stored in WM beforehand, so ordinality stems from LTM (although they may be moved to WM during task execution). However, in other situations, apart from ordinality in LTM (i.e., canonical orders), ordinality can also exist in WM (i.e., orders of items relevant to the current situation or newly memorized orders). In this case, the question arises as to which kind of ordinality is more decisive for spatial mental representations. Previous research has shown that ordinality from WM does not always fully overrule ordinality from LTM and that OPE based on both might coexist (Ginsburg et al., 2014; Ginsburg & Gevers, 2015; Huber et al., 2016; Koch et al., 2023; Zhang et al., 2025). We assumed that this might depend on the LTM and WM salience in the given situation.

WM salience can be systematically manipulated in experimental paradigms, for instance, by asking participants to memorize a new ordinal sequence just before the

task. WM salience can be increased by asking participants to retrieve the sequence in each experimental trial within the task, as compared to only asking them to retrieve it in a memory test after the task. Ginsburg et al. (2014; Ginsburg & Gevers, 2015) did so by using a pure go design, where participants were instructed to respond to all stimuli, versus a go/no-go design, where they were instructed to respond only to stimuli that were part of the learned sequence. In the pure go design, WM content merely needed to be maintained during the task (referred to as intermediate WM salience in Study 4), whereas in the go/no-go design, it needed to be retrieved in each trial (referred to as high WM salience in Study 4). Interestingly, the authors found a SNARC effect in the pure go design and an OPE in the go/no-go design. However, as mentioned earlier, the debate about whether ordinality or cardinality drives the SNARC effect is still ongoing, and in these previous experiments, stimuli from 1 to 10 were used to ensure that ordinality in LTM and cardinality perfectly corresponded to one another. The literature still lacked a systematic investigation into the conditions under which OPE are based on LTM or on WM, which motivated Study 4. Because letters entail LTM-driven ordinality but no cardinality, and because their spatial associations seem to be processed in brain areas similar to those for numbers (Shen et al., 2023), they are well suited for this purpose. Furthermore, OPE have been demonstrated for this kind of stimuli (Gevers et al., 2003); hence, we used letters in Study 4. Note, however, another important difference between letters and numbers beyond whether they entail cardinality: While the Latin alphabet (with some special letters in some languages) contains only 26 letters, there are infinitely many numbers. This means that A and Z will always be the first and last letters in the alphabetical order, whereas for numbers, regardless of which is the smallest or largest in an experimental stimulus set, there will always be smaller and larger numbers outside the experiment. However, researchers mostly use single-digit numbers for investigations of the SNARC effect, and some numerical-cognition effects cannot be generalized from single-digit numbers to multi-digit numbers (Nuerk, Moeller, et al., 2015). When considering only single-digit numbers, the number of possible stimulus sets is limited.

Within single-digit numbers, researchers have investigated stimulus set variations, such as in the seminal SNARC studies by Dehaene et al. (1993) and Fias et al. (1996). Although the question of whether ordinality or cardinality determines the SNARC effect had not yet clearly arisen at the time, these experiments are often cited in the respective discussions. The authors used two different stimulus sets in the typical

PJ task with Arabic digits, namely 0 to 5 versus 4 to 9, to figure out how flexible the SNARC effect was. Data from both studies revealed that the smallest number in each set was associated with the left side, whereas the largest number in each set was associated with the right side. In other words, spatial mappings for the critical numbers 4 and 5, which appeared in both ranges depended on their RM rather than their AM. On the one hand, these results align with the WM account, as the currently used stimulus set is likely represented in WM during task execution. The RM of each number compared to others within this set can be considered its ordinal position, which appears to underlie the spatial mapping. On the other hand, the results are also in accordance with the MNL account, as the currently used stimulus set can be thought of as a part of the MNL being zoomed in to (as described by Pinhas et al., 2013, and Restle, 1970). The findings therefore demonstrate high flexibility of SNA, showing that they adjust to the numerical context, which is consistent with their well-demonstrated dependence on the cultural context, primed spatial mappings, and task demands (Bächtold et al., 1998; Fischer et al., 2010; Lenoir & Badets, 2025; Pfister et al., 2013; Shaki & Fischer, 2008). In a similar vein, SNARC-like effects for physical size have recently been shown to depend on RM instead of AM (Wühr & Richter, 2022). The assumption that SNA depend on RM instead of AM has been cited as a crucial basis for theoretical accounts (e.g., van Dijck et al., 2015), as well as for determining whether an observed phenomenon is actually the SNARC effect (e.g., Di Giorgio et al., 2019; Giurfa et al., 2022; Rugani et al., 2015).

Importantly, results reported in the seminal articles (Dehaene et al., 1993; Fias et al., 1996) seemingly demonstrate the dependence of SNA on RM rather than AM. However, they only consider the absolute number locations on the MNL, and the SNARC effect size was descriptively larger in the low number range than in the high number range in both studies, additionally suggesting a dependence on AM. The null findings in the respective analyses might have been caused by a statistically underpowered design for detecting reasonably small differences between ranges due to the very small sample sizes. Another methodological limitation of both seminal studies is the use of a stimulus set that was not able to rule out several confounds. Specifically, the low range included the number 0, which typically takes longer to be processed (Brysbaert, 1995) and may be represented differently in human minds compared to other digits (Nieder, 2016; Nuerk et al., 2004).

To conclude, the flexibility of SNA in terms of their dependence on RM versus AM is crucial for the research field. We therefore investigated this issue in Study 5 using the typical PJ task with stimulus sets consisting of different ranges of Arabic digits. Note, however, that the dependence of SNA on RM and AM is not mutually exclusive; instead, coexistence is possible.

1.3.4. The SNARC effect's inter- and intraindividual variability

The findings described above show that the SNARC effect arises across different tasks, cultures, and stimuli, suggesting that the phenomenon is very robust. Nevertheless, SNARC research typically follows the tradition of experimental psychology, in which data are averaged across individuals, meaning that effects are examined only at the group level. Recently, more and more researchers have pointed out this "group-to-person generalizability problem" and have developed approaches to investigate the prevalence and size of effects at the individual level (Cipora, van Dijck, et al., 2019; Haaf & Rouder, 2019; McManus et al., 2023; Rouder & Haaf, 2018; Zayas et al., 2019). Importantly, high replicability of a group-level effect does not imply that true underlying scores resemble the group-level effect for all individuals within the group. While this assumption is plausible for phenomena such as the Stroop effect (Stroop, 1935), where all participants are expected to perform better in the congruent than in the incongruent condition, it is not plausible for other phenomena such as beverage preference, where most but not all participants might prefer the sweeter beverage. The first can therefore be called "dominant," whereas the latter are called "indominant" (Rouder & Haaf, 2018). Regarding the SNARC effect, its high replicability at the group level does not necessarily indicate mental representations of numbers mapping small onto the left and large onto the right in all individuals. Rather, this high replicability may reflect that the majority of individuals' mental representations follow the typical mapping (i.e., association of small or large with the left or right, respectively), and only a minority exhibits the reverse pattern. The SNARC effect is an "indominant" phenomenon according to a re-analysis of 18 existing datasets for PJ, where confidence intervals based on individual RT variability were computed for each participant separately (Cipora, van Dijck, et al., 2019). Confidence intervals were computed with three different approaches, and results were similar (psychometric, H0, and H1 bootstrapping). Specifically, while the SNARC effect was observed at the group level, the proportion of participants exhibiting the effect in PJ was only about 45% of

over 1,000 participants. Similar proportions were found for MC (Hohol et al., 2020, Supplementary Materials).

Besides generalizing from groups to individuals, as outlined above, experimental psychologists also seem to implicitly generalize from observations made in single sessions to the typical behavior of individuals. The most common way to investigate whether this is valid is to test participants twice and calculate the test-retest reliability. This measure quantifies how consistent the order of individual scores is between sessions or how stable individual scores are in relation to the average score across individuals (Zorowitz & Niv, 2023). As theoretically explained and experimentally demonstrated by Hedge et al. (2018), effects in experimental psychology can be robust at the group level across sessions, although their test-retest reliability is low. This is in fact likely, because low between-subject variability increases replicability at the group level but leads to low correlations of individual scores between sessions and with other variables. The reason for this "reliability paradox" is that effects are considered in absolute terms when looking at group-level robustness and replicability, while the ranking of individual-level effects within the group is considered when estimating reliability and correlations. High group-level robustness despite low test-retest reliability has, for instance, been found for several cognitive-control measures (Schuch et al., 2022). Reliability may be low, especially for effects that base on difference scores (Caruso, 2006; Cronbach & Furby, 1970), such as the Stroop effect. To calculate the SNARC effect, difference scores between the left and right response sides are calculated as well, and they are additionally even regressed onto number magnitude. As outlined by Hedge et al. (2018), one reason for low reliability of difference scores is the propagation of error from two correlated measures to the difference score. Another reason is that subtracting two measures reduces between-participant variance, thereby increasing the proportion of measurement error in the total variance. In fact, the test-retest reliability of the SNARC effect in PJ appears to be rather small or moderate at best (with estimates of r = .36 by Cipora & Göbel, 2013; r = .41 by Georges et al., 2013; r = .29 by Manavusak, 2025; and r = .27 by Viarouge et al., 2014).

To check the validity of generalizations from single sessions to the typical behavior of individuals, researchers should not only consider temporal stability in *relative* terms, as quantified by test-retest reliability, but also temporal stability in *absolute* terms (Luck, 2019). Importantly, even perfect test-retest reliability for a

phenomenon (i.e., a correlation of individual scores equal to 1 between the test and retest sessions) would not guarantee its temporal stability in absolute terms. This would be the case when all individual scores changed by the same value between sessions. The temporal stability of the SNARC effect in absolute terms remains unclear, although it is crucial for the question of whether the spatial mental representation of number magnitude is a state or a trait.

To summarize, the prevalence and test-retest reliability of the SNARC effect appear lower than one would expect from its high robustness when investigated at the group level. The prevalence has so far not been tested in tasks requiring only judgments of non-semantic number features, so we calculated it for CJ in Study 1. Moreover, intraindividual stability of the SNARC effect remained uninvestigated; we therefore examined it using an extensive longitudinal design in Study 6.

1.4. Internet-based data collection

The data for almost all studies included in this dissertation (Studies 1, 2, 3, 4, and 5) were collected online via the Internet. In the following, I will therefore discuss the advantages of online studies compared to lab studies, as well as the challenges to be tackled when running studies online.

Online experiments provide great opportunities that traditional lab studies cannot offer. First and foremost, they allow researchers to collect large amounts of data efficiently, thereby increasing statistical power (Reips, 2000, 2002). Additionally, compared to lab studies, which often rely on university student samples, online experiments can be run with more diverse samples, thus yielding more heterogeneous and representative datasets. This, in turn, leads to more generalizable conclusions (Reips, 2002; Simons et al., 2017), which, among other reasons, is important for counteracting the replication crisis in the field of psychology. Notably, the SNARC effect was discovered in a lab study (Dehaene et al., 1993), but could be replicated in many online studies (e.g., Cipora, Soltanlou, et al., 2019; Gökaydin et al., 2018; Koch et al., 2023). Our own studies (i.e., Studies 1, 2, 3, and 5 of this dissertation) also demonstrate that the SNARC effect can be found in online experiments.

At the same time, researchers also face methodological challenges when running online studies, which we carefully addressed in our studies or even tried to turn into opportunities. Firstly, response-time measurements seem to be more variable and thus less precise on average in online studies as compared to lab studies, which might be due to differing operating systems or web browsers (Bridges et al., 2020;

Garaizar & Reips, 2019). However, although response-time measurements are the basis of all six studies presented here, the slightly increased imprecision only affects our recordings in an unsystematic way. That is, each study incorporated a large number of trials with many stimulus repetitions with a fully randomized presentation order, and we later excluded outliers. Hence, measurement imprecision should not have distorted any of the reported results and instead may even have increased external validity compared with lab studies (Reips, 2002, 2021).

Secondly, online experiments might catch the attention of Internet users who do not want to participate seriously but instead only want to have a look. This might especially happen when studies are shared publicly and spread across different digital platforms for recruitment purposes. Hence, a seriousness check asking participants about their intentions is crucial in online experiments (Aust et al., 2013; Reips, 2007, 2009). We therefore implemented a question about the participants' intentions at the beginning of each experiment (Studies 1, 2, 3, 4, and 5), and later filtered our data by this criterion: We only ever analyzed data from participants who indicated they wanted to participate seriously, excluding those from participants who indicated they just wanted to have a look at the study.

Thirdly, it is harder to verify information provided by participants in online settings than in lab settings. Although it is safer to rely on objective criteria rather than self-reports (Reips, 2021), even popular recruitment platforms such as Prolific (which we used in Studies 1, 3, and 5) do not offer validation of typical screening criteria, such as age, nationality, or country of residence. Instead, such important demographic information mostly comes from self-reports, even if Prolific verifies each participant's ID during signup to make sure they are a real person with a unique account. We therefore additionally collected demographic information in questionnaires.

Fourthly, the range of devices used for participation in online studies is larger than in lab studies (Reips, 2000, 2021). For example, different browser types, operating systems, and keyboards may be used (Reips, 2007). On the one hand, this can cause unwanted variability in the datasets; on the other hand, this makes the data more representative and thus generalizable. Furthermore, we even used this recorded information to check for multiple submissions by the same participants (as recommended by Birnbaum, 2004, and Reips, 2002). Recognizing that participants potentially use different kinds of keyboards, in several of our studies (Studies 1, 2, 3, and 5), we let participants choose their left and right response keys to ensure their

compliance with our instructions regardless of keyboard type. Although most stayed with the default keys that are in the same row on standard QWERTZ or QWERTY keyboards and about one hand-width apart (i.e., D and K), some participants chose other keys, which might have contributed to our data in a beneficial way.

Finally, the software used by researchers to program their experiments should implement response-time measurements on the client side to ensure higher measurement precision that is mostly unaffected by technical delays. It should also prevent search engine bots from indexing experiment pages and easily proceeding from one to the next, which helps avoid invalid data submissions. To accomplish this, experiments can be equipped with a standardized “noindex, nofollow” meta tag (Reips, 2007). These methodological standards are implemented in WEXTOR, a software that permits researchers to design experiments online in a guided, step-by-step process and to host them afterwards (https://wextor.eu; Reips & Neuhaus, 2002). Because WEXTOR is suitable for both between- and within-subjects designs, we employed it in several studies included in this dissertation (Studies 1, 2, 3, and 5).

Having outlined how the Internet can support data collection and what scientists should consider when making use of it, I will next describe Open Science practices implemented in our studies.

1.5. Open Science practices

To improve the quality of research and accelerate the distribution of new insights, it is crucial that scientists share key parts of the research lifecycle with other scientists. The current “Open Science” movement aims at achieving this by making research more transparent, accessible, and thus reproducible (Crüwell et al., 2019; Spellman et al., 2017; Vicente-Saez & Martinez-Fuentes, 2018). The Internet plays a crucial role in this movement, enabling researchers around the globe to publicly share materials such as stimuli, procedures, data, and analyses. Common Open Science practices include preregistrations, Registered Reports, open materials, open code, open data, open peer review, preprints, and open access publications. Such practices are especially important for fields like psychology that suffer from a replication crisis (Maxwell et al., 2015)[5].

[5] Note that, although generally worth striving for and becoming mainstream, Open Science practices can contribute to structural inequalities (Ross-Hellauer, 2022). For instance, publication fees for open access are typically high and therefore mostly affordable by researchers working at wealthy institutions.

In our studies, we therefore implemented Open Science practices whenever feasible for us (for an overview, see Table 1-1). For instance, we uploaded preregistrations (Studies 1, 3, 4, and 6), which enhance transparency about experimental methods, derivation of hypotheses, and planned data analyses, or even wrote Registered Reports, which additionally include peer-review prior to data collection (Studies 2 and 5). Moreover, we shared study materials necessary for reproducing our research (all studies), such as stimuli, questionnaires, details on experimental procedures, and comprehensive RMarkdown files for statistical power simulations. We also shared our R scripts for data analysis and the collected datasets (all studies). For the Registered Reports, even the peer-review process is publicly available (Studies 2 and 5). Most of our articles are open-access publications (Studies 2, 3, 4, 5, and 6), and their content was made publicly available as preprints even before publication in scientific journals (all studies). By making use of these Open Science practices, we ensured that our research is highly transparent, accessible, and easily reproducible.

Table 1-1

Open Science practices applied in each study with links to publicly available materials

	Study					
Open Science practice	**1**	**2**	**3**	**4**	**5**	**6**
Preregistration	E1, E2		Yes	Yes		Yes
Registered Report		Yes			Yes	
Open materials	E1, E2	Yes	Yes	Yes	Yes	Yes
Open code	E1, E2	Yes	Yes	Yes	Yes	Yes
Open data	E1, E2	Yes	Yes	Yes	Yes	Yes
Open peer review	No	S1	No	No	S1, S2	No
Preprint	Yes	Planned	Yes	Yes	Yes	Yes
Open access publication	Green	Planned	Gold	Gold	Gold	Gold

Note. Abbreviations: E = Experiment, S = stage of RR, Green = green open access (preprint that is freely available in a repository and identical with the published version), Gold = gold open access (publication that is freely available on the publisher's website). Preregistrations in this dissertation refer to non-peer-reviewed preregistrations and are mutually exclusive with Registered Reports.

1.6. Research aims

A large body of past studies has demonstrated that the mental representations of numbers are spatial, as shown by the SNARC effect. However, several controversies in the field of numerical cognition exist to date. Open questions include: How automatic is the SNARC effect truly? In what tasks does it arise? Can it have different shapes? Do culture and language backgrounds play a role? Are similar effects observed in non-numerical ordinal stimuli based on characteristics stored in WM or LTM? To what extent can the SNARC effect flexibly adapt to the context? How inter- and intraindividually stable is the SNARC effect? As numbers are highly relevant in everyday life, it is crucial for cognitive psychology to shed light on these open questions.

To resolve several important controversies surrounding SNA, the present dissertation builds upon six large studies, for which full manuscripts can be found in the Appendix. Study 1 (Appendix 1) aimed to determine how automatic the semantic processing of number magnitude is, as reflected by a potential SNARC effect, by using two CJ tasks that did not require semantic processing of number magnitude. Study 2 (Appendix 2) was a detailed replication of the two most commonly used semantic tasks to measure the SNARC effect, namely, MC and PJ. Study 3 (Appendix 3) was a cross-cultural investigation of the SNARC effect considering script direction as well as other cultural directional preferences, using the same two tasks as in Study 2. Study 4 (Appendix 4) tested whether an OPE for letter ordinality exists in consonant-vowel classification (CVC) and, if so, whether the spatial representation relies on LTM (i.e., the alphabetical order) or WM (i.e., an arbitrary newly memorized order). Study 5 (Appendix 5) focused on the role of RM and AM for the SNARC effect by using variations within one of the most common semantic tasks (i.e., PJ) and within the most common stimuli (i.e., Arabic digits). Finally, Study 6 (Appendix 6) aimed at determining the intraindividual temporal stability of the SNARC effect and other numerical cognition phenomena by using an innovative longitudinal design. A graphical overview of the focus in each study is given in Figure 1-1, whereas Figure 1-2 provides a graphical overview of the tasks used in each study. Further, a summary of the experimental methodology and results in each study is provided in Table 1-2.

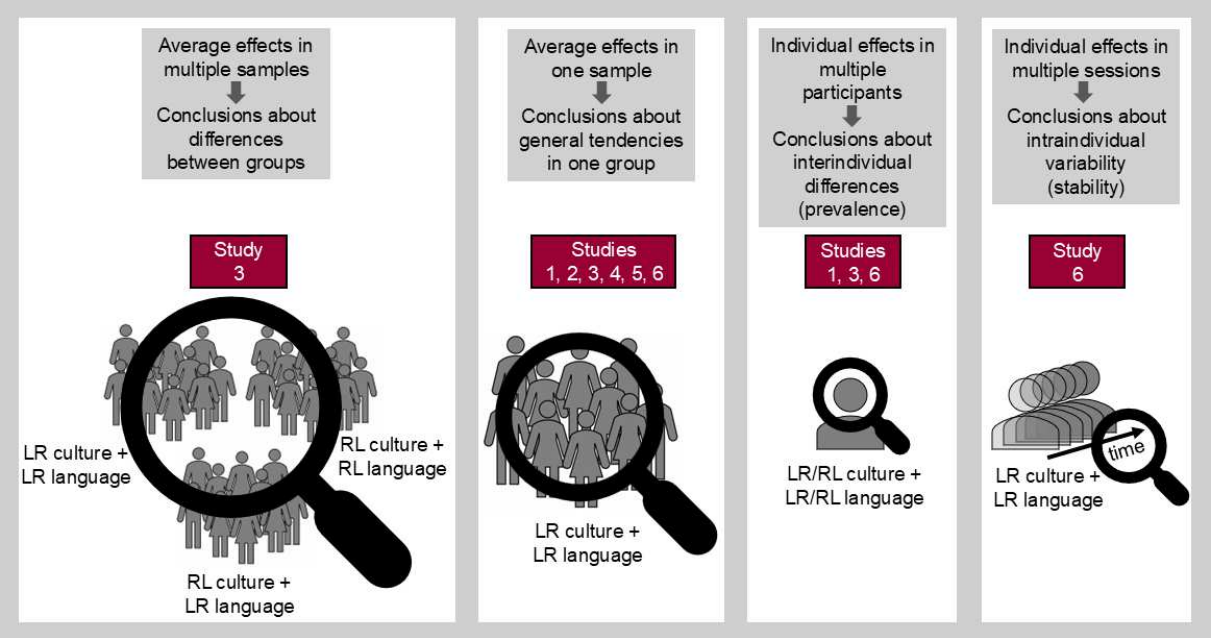
For the studies outlined above, established experimental paradigms from old studies were methodologically improved (e.g., sample sizes and the number of stimulus repetitions per response assignment were optimized) and adapted for online

settings. This permitted us to replicate well-known effects in large samples and investigate them in much more detail. Alongside this, novel paradigms and study designs were created to enable a deeper dive into important questions that the existing literature has not yet been able to resolve. Five out of six studies were run online. This research not only provided additional data, supporting existing claims about the mental representation of number magnitude, but also helped rectify some of them. Taken together, these six studies offer new insights for a deeper understanding of SNA.

Finally, I newly introduce the computational contextualization model that my collaborators and I are currently developing. It reconciles ideas from several existing theoretical accounts by assuming that a combination of cognitive context schemata drives the SNARC and SNARC-like effects. Initial data simulations and parameter optimization results are reported and interpreted.

Figure 1-1

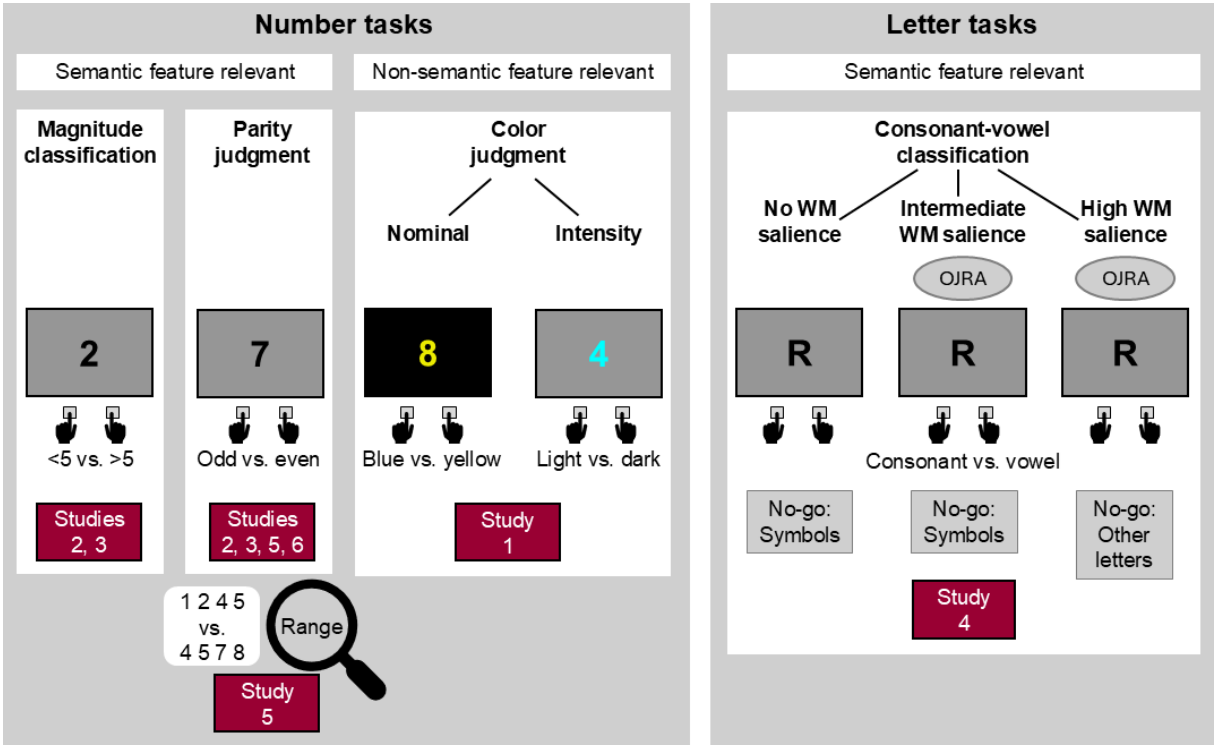
Graphical overview of the focus of investigation in each study



Note. This figure was created in PowerPoint (Microsoft Corporation, 2024). It illustrates the level of investigation with a magnifying glass (multiple samples, one sample, participants, or sessions) and lists what conclusions can be drawn from the data collected at each level. Red text fields indicate in which studies the respective level was investigated, and additional text specifies language and culture characteristics of tested participants.

Figure 1-2

Graphical overview of the tasks used in each study



Note. This figure was created in PowerPoint (Microsoft Corporation, 2024). It provides an overview of all number tasks (left side) and letter tasks (right side). Task names are stated above an illustration of the experimental paradigm, with one example trial each. The decision to be made is explained below the bimanual response-key setup within these illustrations, and red text fields at the bottom indicate in which studies the tasks were used. For PJ, the figure shows that Study 5 looked at different number ranges specifically, and for CVC, the figure shows example sequences to be memorized above the illustrations, if applicable.

Table 1-2

Overview of methods and results of all studies included in this dissertation

Study (Exp.)	Investigated aspect	Sample(s) tested	Stimuli	Task(s)	SNARC (or OPE) slopes
1 (1)	The SNARC effect in non-semantic tasks: Automaticity	360 Western adults (native English, German, Italian, Polish, Portuguese, and Spanish speakers)	Blue vs. yellow Arabic digits from 1 to 9 (excluding 5)	Nominal CJ	−1.71 ms*
1 (2)		360 Western adults (native English, German, Italian, Polish, Portuguese, and Spanish speakers)	Dark blue vs. light blue Arabic digits from 1 to 9 (excluding 5)	Intensity CJ	−1.13 ms*
2	The SNARC effect in common semantic tasks: Shape	701 adult native English speakers	Black Arabic digits from 1 to 9 (excluding 5)	MC, PJ	−20.44 ms*, −3.59 ms*
3	The SNARC effect in different cultures and languages	146 adult native German speakers	Black Arabic digits from 1 to 9 (excluding 5)	MC, PJ	−32.58 ms*, −6.06 ms*
		137 adult native Turkish speakers			−27.26 ms*, −3.43 ms*

		129 adult native Farsi speakers	Black Eastern Arabic digits from 1 to 9 (excluding 5)		−19.42 ms, −1.69 ms*
4 (1)	OPE in letters based on LTM versus WM ordinality	178 Western adults (native German and English speakers)	Black Latin letters E, L, U, W, or A, J, O, R, and symbols ‡, △, ≡, and ∇	CVC without WM sequence (go: letters, no-go: symbols)	LTM: −0.42* (standardized)
4 (2)		169 Western adults (native German and English speakers)		CVC with WM sequence (go: letters in WM sequence, no-go: symbols)	LTM: −0.40* (standardized), WM: −0.08 (standardized)
4 (3)		168 Western adults (native German and English speakers)	Black Latin letters E, L, U, W, and A, J, O, R	CVC with WM sequence (go: letters in WM sequence, no-go: other letters)	LTM: −0.28* (standardized), WM: −0.27* (standardized)
5 (1)	The SNARC effect in different	200 adult native English speakers	Black Arabic digits from	PJ	−5.53 ms* vs. −5.84 ms*

	number ranges:		0 to 5 vs. from 4 to 9		
5 (2)	Flexibility	300 adult native English speakers	Black Arabic digits from 1 to 5 (excluding 3) vs. from 4 to 8 (excluding 6)		−10.17 ms* vs. −5.92 ms*
6	The SNARC effect's inter- and intraindividual variability	10 Western adults (native Dutch, German, and Italian speakers), 30 sessions each	Black Arabic digits from 1 to 9 (excluding 5)	PJ	ranging from −4.83 ms to 0.73 ms (averaged across all sessions per participant)

Note. Sample sizes refer to the number of recruited participants before the application of exclusion criteria. The order of slopes in the rightmost column corresponds to the order of tasks in the column to their left. The slopes for MC in Studies 2 and 3 correspond to categorical predictors, all other slopes correspond to continuous predictors. In Study 4, LTM-based slopes refer to OPE based on the alphabetical order, and WM-based slopes refer to OPE based on the newly memorized letter sequence. Asterisks indicate evidence for slopes different from 0 at the group level (with $p < .05$ in frequentist or Bayes Factors > 3 in Bayesian analyses, and Bonferroni-Holm correction for five tests with a minimal $p < .01$ in Study 4).

2. Results

In the following, I will first explain the statistical analysis typically used to quantify and test the SNARC effect. Then, for the reader's convenience, I will briefly summarize the results from all six studies included in this dissertation, which can also be found in Table 1-2. The full results can be found in the manuscripts in the Appendix.

2.1. Quantification of the SNARC effect

To summarize the key findings of each study, I will first explain how the SNARC and SNARC-like effects are typically quantified. First, mean RT per stimulus (i.e., per number in Studies 1, 2, 3, 5, and 6, and per letter in Study 4) are calculated separately per participant and per response side. Next, differences in mean RT between response sides are computed per stimulus and per participant. These differences, commonly referred to as "dRT" (RT right hand minus RT left hand), are then regressed onto the magnitudes or ordinal positions of the stimuli. A negative regression slope (referred to as "SNARC slope" in the following) reflects an increasing right- relative to left-side advantage with increasing magnitude or ordinal position. Instead of using exact magnitudes or ordinal positions as regression predictors, as done when computing a continuous SNARC effect, magnitudes or ordinal positions can also be categorized (e.g., by contrast-coding numbers 1 to 4 with -0.5 and numbers 6 to 9 with 0.5 in MC) to compute a stepwise SNARC effect. A negative categorical SNARC slope reflects a right- relative to left-side advantage for large compared to small magnitude or early compared to late ordinal position. A more negative SNARC slope reflects a stronger SNARC effect in the typical direction (i.e., association of small/large or early/late with the left/right side, respectively), no matter whether continuous or categorical are considered. Instead of using these unstandardized SNARC slopes, measured in milliseconds, one can also use Fisher-z-transformed standardized SNARC slopes. These can be interpreted as individual SNARC "effect sizes" and are unaffected by the participants' individual means and variances (Cipora, van Dijck, et al., 2019; Tzelgov et al., 2013). As first applied to the SNARC effect by Fias et al. (1996), a repeated-measures regression (Lorch & Myers, 1990) is then computed to test for the presence of the SNARC effect at the sample level: The obtained SNARC slopes are compared with 0 in a *t*-test, and a significant negative effect reflects an overall SNARC effect across all participants.

2.2. Short summary of the main results in each study

In the following, the main results of each study are briefly summarized.

2.2.1. Study 1: High SNARC automaticity shown in non-semantic tasks

To find out to what degree the SNARC effect is automatic, we ran Study 1 with tasks requiring only processing of superficial visual features instead of semantic features of the presented numbers. Following the application of all exclusion criteria as preregistered, datasets from 340 participants for nominal CJ (cyan vs. yellow) and 312 participants for intensity CJ (light cyan vs. dark cyan) remained for analysis, respectively. The data revealed a SNARC effect in both tasks, although small in size (unstandardized slopes of −1.71 ms in nominal CJ and −1.13 ms in intensity CJ), which did not differ between tasks. A reliable regular SNARC effect was revealed in 48% of participants in nominal CJ and 41% in intensity CJ, while a reliable reversed SNARC effect was found in 17% in nominal CJ and 19% in intensity CJ. Importantly, the overall SNARC effect in both tasks was determined solely by small numbers, but not large ones. Specifically, when tested separately, slopes were negative only for numbers 1 to 4, but not for numbers 6 to 9. Moreover, we found little evidence for the MARC and Odd effects. This suggests that processing and spatial mapping of both magnitude and parity take place even when not intended but are more highly automatic for magnitude than for parity.

2.2.2. Study 2: Different shapes of the SNARC effect in different semantic tasks

In Study 2, the two most common tasks involving semantic number processing were administered in a large-scale within-subjects online experiment based on a Registered Report approved by the Peer Community In. After exclusions, data from 609 participants for both MC and PJ were analyzed. The data revealed a replication of the SNARC effect with the standard linear regression model in both tasks, with no difference in effect size between MC (unstandardized slope of −4.03 ms) and PJ (unstandardized slope of −3.59 ms). Crucially, a categorical regression model (unstandardized categorical slope of −20.44 ms) described the SNARC effect better than a continuous one according to the main analyses for MC, indicating a stepwise shape of the SNARC effect. While this reflects that small numbers between 1 and 4 had a similar left-hand advantage and large numbers between 6 and 9 had a similar right-hand advantage, this could not be confirmed by Bayesian leave-one-out cross

validation. Moreover, we observed a linear increase in the right- relative to left-hand advantage within small and large numbers separately. In total, results regarding the shape of the SNARC effect in MC were ambiguous, pointing towards both continuous and categorical parts. The categorical boundary seemed to fall between 4 and 6 for about half the sample in MC, and it was unstable both at the interindividual and intraindividual level. In PJ, model fits were generally lower than for MC. The continuous model fit the PJ data better than the categorical model using a boundary between 4 and 6. The proportions of explained variance and the split-half reliability were higher in MC than in PJ. The correlation between the SNARC effects in both tasks was weak but positive. These findings demonstrate that the SNARC effect differs between tasks.

2.2.3. Study 3: Influence of culture and script direction on the strength of the SNARC effect

The cross-cultural Study 3 involved participants from Germany (LR script and predominantly LR habits), Turkey (LR script and potentially more RL habits), and Iran (RL script and predominantly RL habits). To investigate the influence of language and culture on the SNARC effect, we used the same two tasks as in Study 2 (MC and PJ, with Arabic digits for German and Turkish participants, and Eastern Arabic digits for Iranian participants), as well as the Cultural Directional Preferences Questionnaire (CDPQ) we newly created. The CDPQ comprised seven items in which participants named horizontally arranged objects (from LR or RL), imagined objects (facing rightward or leftward), or arranged a time sequence horizontally (in a LR or RL fashion). After applying all preregistered exclusion criteria, data from 130 German, 112 Turkish, and 75 Iranian participants were analyzed. As expected, cultural directional preferences (measured by a sum score between −7 and 7, for which LR and RL responses were coded as −1 and 1, respectively) were most LR in German (M = −3.65), intermediate in Turkish (M = −2.96), and least LR in Iranian participants (M = 0.07). Similarly, Germans showed the strongest LR SNARC effect (unstandardized categorical slopes of −32.58 ms in MC and unstandardized continuous slopes of −6.06 ms in PJ), Turks an intermediate one (−27.26 ms in MC and −3.43 ms in PJ), and Iranians the weakest one (−19.42 ms in MC and −1.69 ms in PJ). The SNARC effects were significant in each sample and task, except the Iranian sample in MC. A reliable regular SNARC effect was revealed in 57%, 46%, and 39% of participants in MC and by 57%, 38%, and 29% of participants in PJ in the German, Turkish, and Iranian samples, respectively. Crucially, when averaging across participants, neither cultural

directional preferences nor the SNARC effect were RL in any sample, not even in the Iranian one. Although the findings suggest that language script direction and also other cultural directional preferences play a role for SNA, they might not be able to explain them entirely.

2.2.4. Study 4: Coexistence of spatial ordinality representations based on working memory and long-term memory

In Study 4, we investigated whether the spatial mapping of ordinality arises from LTM or WM by running three preregistered online experiments. To avoid confounds with cardinality, we used a CVC task with letters as stimuli. After exclusions, data from 149, 105, and 107 participants were analyzed for Experiments 1, 2, and 3, respectively. The experiments differed in two main manipulations: first, whether or not a letter sequence had to be learned before the task and recalled afterwards (only in Experiments 2 and 3, but not in Experiment 1), and second, whether or not letters from the learned sequence had to be retrieved to differentiate go trials from no-go trials (only in Experiment 3, but not in Experiment 2). In other words, apart from preexisting LTM ordinality (i.e., the alphabetical order), no ordinality in WM was induced in Experiment 1 (no-go: symbols). In Experiment 2, ordinality in WM merely needed to be maintained (no-go: symbols). In Experiment 3, it had to be retrieved in every single trial (no-go: letters not in the sequence). We found a mental spatialization of the alphabetical order stored in LTM (i.e., associations of earlier/later letters in the alphabet associated with the left/right, respectively) in all experiments (standardized slopes of −0.42, −0.40, and −0.28 ms in Experiments 1, 2, and 3, respectively). However, we found a mental spatialization of the order newly stored in WM (i.e., associations of earlier/later letters in the memorized letter sequence with the left/right, respectively) only in Experiment 3 (standardized slope of −0.27 ms). These results provide support for a hybrid model and lead to the conclusion that spatial associations of ordinality stored in LTM and WM can coexist.

2.2.5. Study 5: Partly fixed and partly flexible nature of the SNARC effect

In two online replication experiments based on a Registered Report approved by the Peer Community In, Study 5 investigated the flexibility of the SNARC effect. PJ was assessed in two ranges per experiment: Experiment 1 was a direct online replication of seminal SNARC experiments (Dehaene et al., 1993; Fias et al., 1996) and Experiment 2 was a conceptual replication using an improved stimulus set. After

exclusions, data from 173 and 255 participants remained for analysis in Experiments 1 and 2, respectively. Evidence for a SNARC effect was found in all number ranges. Importantly, the same numbers revealed a right-hand advantage when they were the largest in a range but a left-hand advantage when they were the smallest, supporting previously reported dependence on RM. There was no difference in SNARC slopes between ranges in Experiment 1 (unstandardized slope of –5.53 ms in the lower and –5.84 ms in the higher range), further speaking for dependence on RM. By contrast, in Experiment 2, a stronger SNARC effect was observed in the lower range compared to the higher one (unstandardized slope of –10.17 ms in the lower and –5.92 ms in the higher), providing evidence for additional dependence on AM. To conclude, although high flexibility of SNA is often claimed in the literature, mental representations of number magnitude are partly fixed as well.

2.2.6. Study 6: High inter- and intraindividual variability of the SNARC effect

Finally, in Study 6, the presence and temporal stability of the SNARC effect were investigated at the individual level. For this purpose, an extensive longitudinal, preregistered study design was used, which we call the “Ironman paradigm,” in which 10 participants completed PJ on 30 out of 40 consecutive days. The SNARC effect was present at the group level; however, at the individual level, it resulted from reliable effects in only four to five participants (unstandardized slopes calculated with all data per participant across all sessions ranged from –4.83 to 0.73 ms). Notably, a reliable SNARC effect in the majority of sessions was only found in one participant, and only that participant exhibited significant autocorrelations. Overall, the SNARC effect fluctuated substantially between sessions within participants. It decreased over time along with general response speed, but apart from that, the low temporal stability seemed unsystematic: no correlations of the SNARC effect with hours slept in the previous night, tiredness level, time of the day, and stimulant consumption were detected in most participants. The observed inter- and intraindividual variability challenges frequent generalizations of cognitive phenomena from groups to participants and from participants to sessions.

3. Discussion

In this part, the results from all studies will first be integrated, summarizing insights into influences on the SNARC effect, and then critically discussed with respect to their theoretical implications. After that, a computational model of the SNARC effect will be introduced.

3.1. Insights into influences on the SNARC effect

The studies conducted within this dissertation contribute to the ongoing debates about prominent controversies in the field. In the following, I will interpret the findings with respect to different influences on the SNARC effect. Following the order established in the Introduction, I will first discuss the SNARC effect in different tasks, cultures, stimuli, and time points.

3.1.1. The SNARC effect in automatic and intentional processing tasks

All studies presented in this dissertation clearly demonstrated that the processing of magnitude or ordinality and their mental mapping onto space can take place automatically, not only intentionally. More precisely, SNARC effects (Studies 1, 2, 3, 5, and 6) and OPE for letters (Study 4) were observed in every study. These findings reflect the presence of spatial mental representations of magnitude or ordinality, even when the processing of these features was task-irrelevant. Because we conceive of automaticity as a gradual rather than an all-or-none phenomenon (Moors, 2016), the following discussion will delve into the extent to which the processing of magnitude or ordinality, as well as their mental mappings onto space, are automatic. Unless stated otherwise, interpretations are based on averaged results and therefore apply to general tendencies within groups rather than individuals.

The degree of automaticity when processing and spatially representing numbers. To start with, the degree of automaticity in the mental representation of number magnitude or ordinality appeared considerable in MC (Studies 2 and 3). Although the left and right response keys might have primed the horizontal dimension (as was the case in all experiments we conducted), participants were never instructed to associate specific presented stimuli more strongly with one side than the other. Thus, while semantic processing of numbers was intentional because it was task-relevant, mental spatialization took place in an automatic fashion. It is important to note

that most studies in this dissertation using numbers as stimuli are not able to disentangle whether ordinality or magnitude is the precise feature being processed and mapped onto space. Only Study 4, which used letters, and Study 5, which used different number ranges, can provide further insights into this debate, as will be discussed further below.

Even more interesting insights into the degree of automaticity in number processing (apart from the construction of spatial mental representations described above) come from tasks requiring only the semantic processing of number features other than magnitude or ordinality (e.g., PJ in Studies 2, 3, 5, and 6), or from tasks requiring no semantic processing of number features at all (e.g., nominal CJ and intensity CJ in Study 1). Previous research may have failed to detect truly present but small SNARC effects in CJ mainly due to lacking statistical power. We were able to detect even small SNARC effects thanks to relatively large sample sizes. This was achieved with online data collection in Studies 1, 2, 3, 4, and 5, especially by making use of a recruitment platform (https://www.prolific.com/) in Studies 1, 2, and 5, which is very efficient (Reips, 2000, 2002, 2007). The fact that we observed the SNARC effect across all of these tasks indicates that number processing including the activation of mental representations of numbers, is highly automatic (in line with Tzelgov et al., 2015), at least when looking at the group level. Accordingly, previously reported null results from frequentist analyses in CJ may simply reflect insufficient statistical power and should not be interpreted as evidence for the absence of the SNARC effect. Importantly, when number magnitude task-irrelevant, the strength of an observed SNARC effect does not reflect the strength of automatic semantic magnitude processing, nor does it reflect the strength of the spatial association (Cipora, He, et al., 2020). There might be an interplay between both aspects producing the observed SNARC effect.

On a side note, beyond magnitude or ordinality, other semantic features of numbers, such as parity, are sometimes also associated with horizontal directional space. However, in the case of the MARC effect, the evidence was weaker and less consistent (Studies 1, 2, and 3, although more robust in Study 6) as compared to the almost consistently found evidence for the SNARC effect. As a conclusion, the processing and spatial mapping of number magnitude or ordinality seem to be more automatic than those of number parity.

The strength and prevalence of the SNARC effect. Predictions regarding the strength of the SNARC effect in different tasks can be derived from two lines of argumentation: First, the tasks typically differ in overall response speed, and the SNARC effect has been found to be stronger in longer RT (Cipora, Soltanlou, et al., 2019, Supplementary Materials; Didino et al., 2023; Gevers et al., 2010). This would speak for the strongest SNARC effect in PJ (M_{RT} = 577 ms in Study 2), an intermediate one in MC (M_{RT} = 527 ms in Study 2), and the weakest one in CJ (M_{RT} = 460 ms in nominal CJ or M_{RT} = 443 ms in intensity CJ in Study 1). Second, the SNARC effect might be stronger the more magnitude or ordinality, other semantic features, or spatial aspects are task-relevant (see Figure 7 in Study 1). This would predict the strongest SNARC effect in MC, an intermediate one in PJ, and the weakest one in CJ. Possibly reflecting a combination of both predictions, we found similarly sized SNARC effects in MC and PJ (unstandardized continuous slopes of –4.03 ms and –3.59 ms in Study 2, see also Study 3) and a weaker SNARC effect in CJ (–1.71 ms and –1.13 ms in Study 1).

Importantly, although a typical SNARC effect can be demonstrated in MC, PJ, and CJ at the group level, most individuals seem not to reveal it reliably when taking into account intraindividual variability of RT. This limitation was revealed when calculating bootstrap confidence intervals (see Study 1 for CJ; Hohol et al., 2020, Supplementary Materials, for MC; Study 6 and Cipora, van Dijck, et al., 2019, for PJ). Conclusions can therefore only be drawn about average SNARC effects, and generalizations from groups to individuals should be avoided.

A natural follow-up question following up on the above findings is how strong the SNARC effect is in individuals whose semantic number processing is highly automatized, such as professional mathematicians. While no SNARC effect was found in PJ among professional mathematicians (Cipora et al., 2016), no differences were found between professional mathematicians and controls for other numerical-cognition phenomena that typically arise in MC (Hohol et al., 2020). This might either reflect that such phenomena are similar between those groups, or that they differ between groups with several processes working in opposite directions (Cipora, He, et al., 2020). To further investigate the automaticity in number-magnitude processing and its mental spatialization, professional mathematicians are currently being assessed with CJ at the Universities of Krakow, Graz, and Padua in a new collaboration in which I am involved (Cipora et al., 2024).

The shape of the SNARC effect. As described in the Introduction, previous research has already suggested that the SNARC effect does not always look the same. No study, however, has so far examined its shape in the two most common semantic tasks this thoroughly by testing plausible statistical models against each other with such a large dataset from a within-subjects design. As anticipated, we indeed found a rather continuous SNARC effect in PJ, and a more categorical SNARC effect in MC. Nevertheless, the observed stepwise shape of the SNARC effect in MC might not be the whole story. First, a continuous proportion of the SNARC effect seemed apparent upon visual inspection (see Figure 4 in Study 2). Second, the model fit using the continuous predictor was only slightly lower than that using the categorical predictor. Third, Bayesian leave-one-out cross validation indicated the opposite. Finally, there was a significant increase in the right- relative to left-side advantage in responses within numbers 1 to 4 and numbers 6 to 9 when tested separately.

The findings regarding the shape of the SNARC effect in MC are not fully consistent with results reported in the literature. An early meta-analysis showed an almost uncontroversial, purely categorical SNARC effect in MC (see Figure 1 and Table 2 in Wood et al., 2008). By contrast, our own Study 3 seemed to reveal a categorical step in dRT between small and large numbers, but also a negative slope in small numbers, at least at the descriptive level, across all samples. Our results from Study 2 reveal even another different pattern: an overall categorical shape with continuous aspects in both small and large numbers. Taken together, researchers should consider switching from traditionally used continuous to categorical statistical models for the SNARC effect in MC, at least as a first step. If a plot of dRT per continuous magnitude suggests a continuous component of the MC-SNARC like in Study 2, researchers should additionally check in a second step whether continuous slopes computed separately for small and large numbers are negative. Moreover, researchers should consider interindividual variability of the SNARC effect, as the group-level pattern might also originate from a part of the sample revealing a continuous shape and a different part of the sample revealing a categorical shape.

Notably, we also found a mixture of a continuous and categorical SNARC effect in non-semantic number judgments; however, the exact shape differed from that observed for semantic number judgments described above. Specifically, in both nominal and intensity CJ used in Study 1, we found a significant overall SNARC effect. Crucially, the shape was inconsistent across the stimulus set: Only small numbers

were mentally mapped onto space. That is, the relative right- over left-hand advantage increased in the interval from 1 to 3 or 4, whereas numbers in the interval from 6 to 9 showed a similar right-hand advantage. Such a tendency may reflect a logarithmic MNL, as suggested in studies using not only single-digit numbers but also multi-digit numbers (Dehaene, 2003; Nuerk, Moeller, et al., 2015). Interestingly, when looking at dRT for small and large numbers separately in MC and PJ (see Figures 4 and 5 in Study 2, and Figure 3 in Study 3), the decrease appears descriptively slightly steeper for numbers 1 to 4 than for numbers 6 to 9 in all samples. This is in line with our results from Study 5, discussed further below, in which participants were assessed with small and large numbers separately. To summarize, the mental representation of small numbers seems to be more automatized, at least in the single-digit range.

Conclusion. The collective findings from all mentioned studies demonstrate that the processing of magnitude or ordinality can either take place intentionally (i.e., when required by task instructions), or automatically (i.e., when irrelevant to the task at hand). This could be inferred from the highly automatic spatial mental representations of numbers, which, although not always consistent in shape, were robustly revealed – at least when averaging across participants.

3.1.2. The SNARC effect in different cultures and languages

As outlined in the Introduction, the direction of the SNARC effect has been suggested to align with the script direction of a language (most prominently in the MNL account). That is, small numbers are said to be associated with the left and large numbers with the right in LR languages, whereas the reverse is expected in RL languages. However, this might not be the full story.

Script direction and its relation to the SNARC effect. The hypothesis of a link between script direction and the direction of the SNARC effect has been supported by early comparisons of samples with lifelong exposure to LR versus RL scripts (Shaki et al., 2009; Zebian, 2005), by investigations with migrants who transitioned from RL to LR script environments (e.g., Dehaene et al., 1993), and by explorations of situational influences in bilinguals, such as priming LR or RL contexts (Fischer et al., 2009; Shaki & Fischer, 2008). However, more recent studies have failed to replicate a reversed SNARC effect in samples using a RL language (Cipora, van Dijck, et al., 2019; Feldman et al., 2019; Hochman et al., 2024; Lopiccolo & Chang, 2021; Zohar-Shai et al., 2017). In Study 3, we found regular SNARC effects in each of the samples and tasks at the group level, except for the MC task in Iranians. Crucially, Iranian

participants did not show a reversed SNARC effect, as might have been predicted considering their RL script. These findings suggest that script direction alone might not be the only factor influencing the strength and direction of the SNARC effect.

Cultural directional preferences and their relation to the SNARC effect. Beyond script direction, other directional preferences may shape the SNARC effect, potentially emerging as early as in preliterate childhood (Patro, Nuerk, et al., 2016). Although cultural directional preferences often align with script direction (Faghihi & Vaid, 2023), this is not necessarily true for every culture. In particular, Turkish cultural directional preferences might not be as strongly LR as those of other cultures with a LR script. This can be attributed to a historical shift from a RL to a LR script around a century ago, coupled with RL practices encouraged by Islamic traditions. The CDPQ, newly developed for Study 3, supported this prediction. Directional preferences were only modestly LR in the Turkish sample, positioned between the clearly LR German and mixed LR and RL Iranian samples, without any strong directional tendency. This pattern was mirrored in the group-level SNARC effects, which were strongest in the German sample, intermediate in the Turkish sample, and weakest in the Iranian sample, in both PJ and MC. Note that the CDPQ scores and SNARC effects reported above refer to averaged results across participants. At the individual level, no correlations were detected between the CDPQ score and the SNARC effects; however, group-level cultural directional preferences mirrored SNARC effects, consistent with our prediction. Null correlations may be due to the high intraindividual variability of the SNARC effect, as shown in Study 6, which undermines their reliability and makes the SNARC effect largely unsuitable for correlations with other measures.

The CDPQ as a diagnostic tool. Overall, the CDPQ appears to be a promising tool for successfully capturing cultural directional preferences. However, as it was used for the very first time in Study 3, several refinements are advisable. First of all, items with low discrimination power between samples (e.g., Item 4: Bike liking) could be omitted, or items could be weighted unequally depending on their discrimination power. Second, items related to directional behavior recommended in the Qur'an, the holy book of Islam, could be assessed. For instance, the Prophet Muhammad taught that one should eat and drink with the right hand and begin actions such as putting on shoes on the right side (Parrott, 2013, 2017), which might lead to stronger RL cultural directional preferences in Muslim participants. Third, psychometric and diagnostic properties such as construct validity and test-retest reliability need to be systematically

assessed. To figure out how variable CDPQ scores are within individuals, we are currently assessing a revised offline version of the CDPQ (the CDPQ-R) with German-Turkish bilinguals in different cultural contexts (Roth et al., 2025). Specifically, the experimenter only speaks German, and all materials are provided in German in one session, whereas the experimenter only speaks Turkish, and all materials are provided in Turkish in a different session. Once the full sample will be collected, we plan to compute the correlation of CDPQ-R scores between cultural contexts. Fourth, establishing normative data from several different larger samples is essential. These samples should be carefully selected to represent different script directions, regions of residence, and other cultural background factors. Finally, although the CDPQ was designed for online administration, developing an alternative version for lab studies could also prove useful. While less efficient, an offline version of the CDPQ would allow for more interactive assessments and could thereby potentially capture embodied preferences and spontaneous behavior more realistically. We implemented this for the new study testing German-Turkish bilinguals mentioned above. For instance, we replaced questions on how participants would hypothetically draw or arrange objects with new items requiring them to actually perform these actions, with the experimenter coding their behavior as predominantly LR or RL.

No evidence for cultural RL preferences and reversed SNARC effects. The finding that both the group-level CDPQ sum scores and the SNARC effect in MC and PJ did not exhibit any RL tendency despite their RL script initially appears surprising. One plausible explanation for this is the globalized, predominantly LR-oriented culture. For instance, globally consumed media (e.g., U.S. American movies) and widely used technologies (e.g., operating systems developed by U.S. American companies) often feature LR directional functionalities and might thus reinforce LR directional preferences. Consequently, even individuals using RL scripts may be influenced by exposure to cultural LR directionality in their everyday lives. It would therefore be interesting to investigate the SNARC effect in relation to cultural directional preferences in remote samples using a RL script. However, finding such samples has become increasingly difficult, as the majority of the world's population has access to globalized culture nowadays (e.g., through the Internet and travel opportunities), and because RL languages are inherently less common than LR languages. Researchers could alternatively recruit from rural communities within countries using RL scripts that have limited exposure to urban multiculturalism and international media, or from

politically isolated regions where access to global technologies and content is restricted. If such samples were to show stronger RL preferences and a reversed SNARC effect, this would support the idea that globalization attenuates the impact of RL directionality in culture, including language script. Conversely, a lack of stronger RL directional preferences and a SNARC effect showing no reverse tendency would indicate that culture and script direction are not the sole factors at play. Instead, the SNARC effect may have a partially innate basis (Rugani & de Hevia, 2017). This would be in line with studies indicating LR SNA in preverbal humans (de Hevia et al., 2014, 2017; di Giorgio et al., 2019) and non-human species (Drucker & Brannon, 2014; Gazes et al., 2017; Giurfa et al., 2022; Rugani et al., 2010, 2015).

Conclusion. To sum up, our findings indicate that while the SNARC effect partly aligns with cultural directional preferences and script direction, these factors alone cannot fully account for its strength and direction. The absence of a reversed SNARC effect even in RL-script samples, at least when averaged across individuals, highlights the necessity of considering a broader array of factors (e.g., historical, cultural, technological, and biological).

3.1.3. The SNARC effect in different stimuli

So far, we have only looked at results from studies using numbers 1 to 9 (excluding 5) as stimuli. However, as outlined in the Introduction, the SNARC and SNARC-like effects such as OPE have been demonstrated in many studies using different kinds of stimuli. In the following, I will first discuss the findings from Study 5, which compared several number ranges used as stimulus sets in PJ and relate these findings to those from previously discussed studies. I will then turn to the findings from Study 4, which investigated the mental representations of letters instead of numbers. Note that Studies 4 and 5 were run from the perspective of general cognition, focusing on group-level tendencies rather than effects at the individual level.

Flexible and fixed SNARC aspects found in different number ranges. When examining and comparing the SNARC effect separately across different number ranges, the mapping of specific numbers onto the MNL in absolute terms is only one aspect to consider, and this has already been done in the seminal studies (Dehaene et al., 1993; Fias et al., 1996). Another important aspect researchers can examine is the strength of the SNARC effect in different number intervals in terms of how steep the negative slope is. In Study 5, we replicated a right-side association of numbers 4 and 5 when they constitute the largest numbers in the current stimulus set, but a left-

side association when they are the smallest (Experiments 1 and 2). This demonstrates that the SNARC effect is at least partly flexible, with the spatial association of each number being influenced by its magnitude relative to the other numbers in the given set. This is in line with the notion of a ranking-space (Abrahamse & van Dijck, 2023). In addition, however, we demonstrated that the strength of the SNARC effect is also influenced by AM (see conceptual replication in Experiment 2). Specifically, the SNARC slope was more negative when the stimulus set included only numbers from 1 to 5 (excluding 3) than when it included only numbers from 4 to 8 (excluding 6). This pattern became observable only after refining the stimulus set compared to the seminal papers (Dehaene et al., 1993; Fias et al., 1996) to avoid potential confounds: In particular, we excluded number 0 because of its special properties (Brysbaert, 1995; Nieder, 2016; Nuerk et al., 2004), and the middle number of each range to prevent potential interactions with the MARC effect (Nuerk et al., 2004). This experiment supports recent studies advocating for the usefulness of conceptual online replications in the field of cognition (Miccoli & Reips, 2025). To summarize, the findings from Study 5 indicate that the SNARC effect is not entirely flexible but also partly fixed. Moreover, they provide further evidence for a more precise or more strongly spatialized mental representation of small numbers compared to large ones, as already shown in Studies 1, 2, and 3.

LTM- and WM-based OPE for letters. Beyond numbers, other kinds of ordinal stimuli (e.g., months, weekdays, letters) exist, and their ordinality is also mentally spatialized. The OPE found in all three experiments of Study 4 using indirect CVC tasks clearly show that the mental spatialization of letter ordinality stored in LTM is also highly automatic: Although the alphabetical order was task-irrelevant in all three experiments, it was still activated and associated with space by participants, at least when considering the average processing at the group level. Moreover, we also found stimulus-specific associations for letters. Namely, the spatial mental representation of letter ordinality stored in LTM seems to be stronger for the letter A than for the other letters we assessed in Study 4. We observed a left-response advantage for letter A, irrespective of its position in a sequence newly stored in WM. In other words, including the letter A in the stimulus set may increase the likelihood that a potential OPE based on WM gets overridden by an OPE based on LTM. In Study 4, we only used the stimulus sets A, J, O, R, and E, L, U, W. It is therefore hard to tell whether the relatively strong spatial association only holds for letter A because it is the first letter of the Latin

alphabet, or whether the OPE based on LTM is stronger for letters early in the alphabet in general. To conclude, ordinality stored in LTM clearly influenced the mental representation of letters.

In addition, ordinality arising from WM also seems to be processed and mapped onto space, as shown by the OPE observed in Experiment 3 in Study 4. However, in contrast to more robust OPE based on the alphabetical order stored in LTM, the OPE based on a sequence newly stored in WM was only elicited when that sequence was made salient. Specifically, in the go/no-go design in Experiment 3, for each stimulus participants were presented with (e.g., E), they had to retrieve the sequence (e.g., LEWU) to determine whether the stimulus was part of it (e.g., E is indeed part of LEWU). However, only the presence of stimuli in the sequence was relevant, whereas their order remained irrelevant, even in Experiment 3.

In a new, yet unpublished study (Albu, Slawinski, et al., 2025), we conducted three similar experiments, with the only difference to Study 4 being that ordinality stored in LTM is task-relevant: Instead of an indirect task, we used a direct one in which participants had to judge whether a letter came before or after K (for the sequences OJRA and RAOJ) or O (for the sequences LEWU and ULEW) in the alphabet. Preliminary results reveal the expected LTM-based OPE in all three experiments. Interestingly, a WM-based OPE was not only observed with high WM salience, but instead also with intermediate WM salience. That is, spatial associations of WM ordinality may persist even with high LTM salience, as long as WM remains salient. In another new, yet unpublished study (Albu, Scheiderer, et al., 2025), we conducted three similar experiments to see whether the results extend to other ordinal stimuli, namely names of months, using an indirect task like in Study 4. Notably, the mental representation of LTM ordinality in months may be even stronger than for letters (i.e., most people can quickly retrieve the preceding and following months for each month in the year, but may not be able to retrieve the preceding and following letter for each letter in the alphabet). Preliminary results reveal a LTM-based OPE when WM salience was low or intermediate, and a WM-based OPE when WM salience was high, but no coexistence of the effects. To conclude, spatial associations of ordinality exist for content stored in LTM or WM, and they arise both in direct and indirect tasks.

An important potential limitation of Study 4 and of our two new studies mentioned above is that only one sequence was memorized per participant, in contrast to studies where participants had to learn many different sequences throughout the

experiment (Guida et al., 2015; van Dijck & Fias, 2011). Sternberg (2016), who calls the first a varied-set procedure and the latter a fixed-set procedure (Sternberg, 1966), argues that sequences in a fixed-set procedure are in fact stored in LTM rather than in WM. Relatedly, a recent study showed that an OPE based on a new sequence stored in WM may disappear during an experiment as the sequence is moved to LTM (Ftaïta et al., 2024). Whether the OPE resulting from the newly learned letter sequence in Study 4 truly reflected a mental representation of ordinality coming from WM or instead also from LTM is therefore uncertain.

Memory scanning processes. An interesting discussion in the literature exists about the nature of memory scanning processes (e.g., judging whether a letter is contained in a newly memorized ordinal sequence). An important distinction is made when it comes to whether they are serial, describing that each item stored in memory is checked separately (Sternberg, 1966), or parallel, describing that all items are scanned simultaneously (Ratcliff, 1978). In case the scanning process is serial, a follow-up question is whether it is exhaustive or self-terminating (for an overview, see Sternberg, 2016). On the one hand, exhaustive search describes that humans scan the entire sequence stored in WM or LTM (Sternberg, 1966), even when the stimulus is already found early in the sequence. Evidence from early studies supporting serial exhaustive search (Sternberg, 1966) implies that humans make unnecessary comparisons (as outlined by Crowder, 1976, p. 357), points towards highly automatic ordinality processing. On the other hand, self-terminating search refers to the memory scanning being stopped as soon as a match has occurred (Sternberg, 1969), which seems more efficient at first sight. However, memory search might be highly automatic, so that additional cognitive effort might be required for its interruption. For instance, if the determination of whether a match has occurred, executing this operation after each comparison (i.e., between the searched item and each item included in the sequence stored in memory) would be less efficient than an exhaustive search (Sternberg, 1966).

Memory scanning is also required in studies investigating the OPE using go/no-go paradigms. While Sternberg's paradigms consist in mere memory scanning, this is only the first step in OPE experiments, in which a binary decision is to be made as a second step if a match was found. Previous papers on OPE investigations reported an increase of average RT in go trials along with the items' ordinal position in the sequence stored in WM (van Dijck et al., 2013; van Dijck & Fias, 2011), pointing towards serial self-terminating memory scanning (Sternberg, 1969). Such a trend is

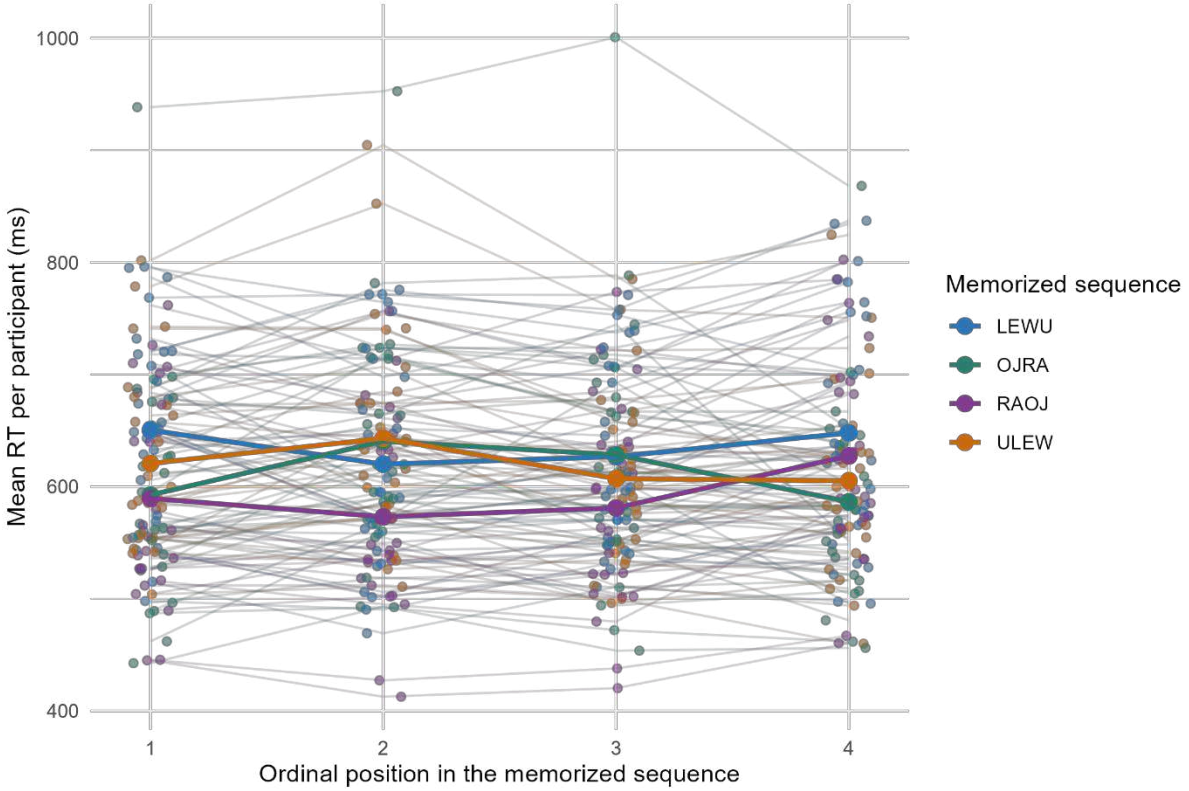
not visible in Figure 3-1, which illustrates the data we collected for Experiment 3 in Study 4. The lack of RT increase in go trials along with ordinal position may suggest serial exhaustive or even parallel memory scanning. Each participant had to memorize and retrieve only one sequence, leading to a highly automatic memory search within that sequence by training across many trials. This, in turn, may have led to serial exhaustive or parallel memory scanning being more efficient than serial self-terminating memory scanning. Alternatively, the fact that we only used two out of 24 possible sequences within each letter set (i.e., E, L, U, W, or A, J, O, R) may have obscured an RT increase and truly underlying serial self-terminating memory scanning, because RT often differ between stimuli per se. For instance, L seemed to be responded to more slowly (see LEWU in blue and ULEW in orange in Figure 3-1) than A (see OJRA in green and RAOJ in pink in Figure 3-1) in the CVC task we used in Experiment 3 in Study 4. To properly distinguish between serial self-terminating and serial exhaustive or parallel memory scanning while avoiding confounds with stimulus-specific RT, it would be necessary to include all 24 possible sequences within each letter set. To conclude, our data suggests serial exhaustive or even parallel memory scanning, although it is not perfectly suitable for contributing to the discussion.

Beyond the debate about serial (exhaustive or self-terminating) versus parallel memory search, items positioned at the beginning or end of a set were suggested to be less subject to so-called proactive or retroactive interference caused by neighbor items than items positioned in the center of a set (Foucault, 1928). Accordingly, the first and last items in a memorized sequence should be more easily retrieved than items in the middle of a sequence (as outlined by Crowder, 1976, p. 446), as empirically reflected in so-called primacy and recency effects (Postman & Phillips, 1965). Similarly, in numerical cognition, the so-called end effect shows faster responses when numbers from the beginning or end of a given sequence are used in number comparison tasks (Lozin & Pinhas, 2024). However, RT in Experiment 3 in Study 4 seem similar for ordinal positions 1 and 4 as for 2 and 3 (see Figure 3-1), thus not providing evidence for primacy or recency effects. This might be due to participants only memorizing one sequence and rehearsing it many times across the experiment, thereby decreasing the advantage of the first and last items as compared to the central ones. Alternatively, primacy and recency effects might be obscured by stimulus-specific RT, as explained in the previous paragraph. To conclude, our data does not

indicate primacy or recency effects; however, they are insufficient to draw firm conclusions.

Figure 3-1

Mean RT per stimulus in go trials for all four memorized sequences used in Experiment 3 in Study 4



Note. This figure was created in R (R Core Team, 2025). It displays the mean RT per stimulus according to their ordinal position in the respective memorized sequence for Experiment 3 in Study 4, where WM salience was high. Different sequences are coded with different colors (e.g., blue data illustrates data for letters L, E, W, and U, corresponding to the ordinal positions 1, 2, 3, and 4, respectively). Smaller points (slightly jittered along the x-axis for better visibility) and thinner connecting lines reflect the mean data per participant separately, while larger points and thicker connecting lines reflect the mean data across all participants.

Conclusion. Taken together, Studies 4 and 5 add to the body of evidence for the SNARC effect and OPE for different stimulus sets, such as various number ranges and letters. Moreover, we were able to show that the nature of these effects is partly flexible, but at the same time also partly fixed, at least when investigated at the group level. These effects can be driven by both magnitude and ordinality stored in LTM, as well as by newly acquired ordinality encoded in WM, operating simultaneously.

3.1.4. The SNARC effect's inter- and intraindividual variability

After having discussed the limited generalizability of the typical SNARC effect from one culture to the next, I will point out now that its generalizability is also limited in other respects, as shown in Study 6. First, generalizability from samples to participants is limited, since a participant's typical behavior is not necessarily similar to the average behavior of a group of participants. Second, generalizability from single experimental sessions to typical behavior is limited, meaning that data collected at one point in time can deviate from data collected from the same participants at different time points.

Generalizing from groups to individuals. Regarding the first restriction on generalizability, results from Studies 1 and 6 suggested that an indominant nature of the SNARC effect (Rouder & Haaf, 2018), characterized by large between-subjects variability. Although we only tested 10 participants in Study 6 using PJ, the prevalence of reliable regular SNARC effects (40% to 50%) and reliable reversed SNARC effects (10%) at the intraindividual level was comparable to findings from an earlier study testing over 1,000 participants using PJ. The prevalence of reliable regular SNARC effects in both CJ tasks (40% to 50%) assessed in Study 1 was similar to that in MC (Hohol et al., 2020, Supplementary Material) and PJ (Cipora, van Dijck, et al., 2019); however, the proportion of the samples (17% to 19%) showing a reliable reversed SNARC effects was larger in CJ than in PJ, and more similar to MC (13% to 23%). In most published SNARC studies, researchers have reported only group-level effects, following the traditions of experimental psychology. Those researchers mainly try to shed light on general tendencies in cognitive processing, where between-subjects variability is considered disruptive and is therefore minimized or ignored by averaging data from many participants. In contrast, researchers working within the traditions of differential psychology examine cognitive effects at the individual level, seeking relations with other variables. Here, between-subject variability is not noise but instead the target of analysis, and researchers attempt to maximize it. Relatedly, researchers should be aware of the "reliability paradox" (Hedge et al., 2018), which describes that low between-subject variability enhances replicability at the group level, but simultaneously goes hand in hand with low intraindividual correlations of effects between time points or effects with other variables. Our findings add to the recently growing body of literature on this issue in cognitive research (e.g., Cipora et al., 2019; Haaf & Rouder, 2019; McManus et al., 2023; Rouder & Haaf, 2018; Zayas et al., 2019).

In a recent preprint, Rouder et al. (2025) even argue that psychological research should report more measures of task goodness than only reliability along with results. To conclude, combining goals from experimental and differential psychology poses challenges. Nevertheless, researchers must carefully choose the paradigms they use according to their goals. For instance, the classic PJ task seems to be more adequate for investigating a group-level rather than individual-level SNARC effect.

Generalizing from single observations to typical behavior. Regarding the second generalizability restriction, the PJ data collected within the extensive longitudinal design of Study 6 suggested large unsystematic fluctuations at the intraindividual level. Although the SNARC effect is a stable group-level phenomenon, having been replicated in countless studies, it seems to emerge from unstable individual-level response-time data (as also found by Hedge et al., 2018 in the case of MC). A possible interpretation is that a distribution of plausible SNARC effects, rather than a stable effect size, underlies the measured response-time data for each individual. Depending on the time point and situational influences (Cipora et al., 2018), individuals may reveal a regular, null, or reversed SNARC effect – but importantly, these outcomes are not equally probable. Note that this idea closely relates to the concept of contextualization introduced for a new computational model in Section 3.3. Beyond the most plausible SNARC effects, such distributions might additionally differ between individuals regarding their shapes and widths. For instance, the central tendency (i.e., mean, median, or mode) of the distribution for the only participant in Study 6 who showed a SNARC effect in the majority of sessions might be a rather strong SNARC effect, and their density for reversed SNARC effects might be small. This distributional explanation would lead to the conclusion that the SNARC effect (or the mental representation of number magnitude reflected in it) simultaneously entails trait- and state-like components: The individual distribution of plausible SNARC effects could be considered a trait, whereas the observable SNARC effect at a specific time point, drawn from that distribution, varies as a function of both the time point and situational influences, and could therefore be considered a state. It should be noted that large, unsystematic fluctuations could also result from methodological flaws causing measurement imprecision, which would invalidate the above-mentioned speculations. However, we believe this is not the case, because the intraindividual stability of RT was higher and varied in a more systematic way, with participants on average becoming faster from session to session.

Temporal stability of the SNARC effect. Although intraindividual variability was considerable, on average, the SNARC effect in PJ diminished over a period of just over a month (Study 6). However, this result should be interpreted cautiously with respect to the development of automatic number-magnitude processing and mental spatialization. At first glance, this finding might suggest that number processing, or its mapping onto space, becomes less automatic over time or with increasing task familiarity. Yet, response speed improved over time, and the SNARC effect is typically weaker in faster responses than in slower ones, as mentioned earlier. Disentangling the strength of mental number spatialization over time from overall response speed in PJ is hardly feasible. For within-subject comparisons, this would require participants to maintain a constant response speed across sessions, but most individuals become faster with practice. For between-subject comparisons, the SNARC effect would need to be more consistent across individuals than it actually is, in terms of its presence, strength, and even direction. To better investigate the degree of automaticity in number processing and mental spatialization over time, future research might consider using non-speeded binary preference paradigms instead of speeded judgment tasks to avoid overall RT as a confounding variable.

Limitations and future research directions. An important limitation of Study 6 is that only 10 participants were tested, some of whom were aware of the research question. Therefore, our sample may not be representative of the broader population, or even of individuals with similar age, gender, culture, and educational backgrounds. Still, only a few studies (e.g., Judd et al., 2023) with similarly extensive longitudinal designs exist in cognitive psychology to date, so the main contribution of Study 6 to the field may lie more in its innovativeness than in its representativeness. As one of the first studies of its kind, it certainly lays the groundwork for further investigations into the intraindividual stability of cognitive effects. Another limitation of the study is that only four potential covariates of the SNARC effect were considered (i.e., hours slept in the previous night, tiredness level, time of the day, and stimulant consumption). While we found no correlations with these, many other covariates could be investigated (e.g., motivational and attentional factors, body-related factors such as appetite or hormone levels, environment-related factors such as noisiness or temperature, priming-related factors such as number context).

As an interim conclusion, future studies should be conducted to assess the SNARC effect in the “Ironman paradigm” using larger samples and different tasks,

while also considering more covariates. Furthermore, it is important for the field to figure out whether only the SNARC effect is highly variable between and within individuals, or whether other cognitive effects show similar variability. In Study 6, we found a MARC effect in six participants and an Odd effect in five participants, suggesting prevalence rates comparable to those of the SNARC effect. When looking at fluctuations over time, the MARC and Odd effects appeared similarly unstable as the SNARC effect, but drawing conclusions from only 10 participants is risky. Going one step further, echoing a recent study of several cognitive tasks with a large sample of children (Judd et al., 2023), another important research question is whether an ubiquitous variability trait characterizes intraindividual variability across tasks. According to that study, intraindividual variability seems to be rather task-specific. However, this remains to be investigated for numerical-cognition effects and in adults.

To try to answer several of these fundamental questions, we are currently conducting a follow-up study with 34 participants completing 25 sessions, each including four tasks with different kinds of outcome variables (Cipora et al., 2023). The tasks include: (i) mere detection of visual stimuli, (ii) magnitude comparison for two symbolic numbers, where performance in (i) and (ii) is quantified by the median RT, (iii) a non-symbolic magnitude comparison task, where the Approximate Number System acuity is quantified by overall accuracy, and (iv) physical size judgments of symbolic numbers, where the Numerical Stroop effect is quantified by the average difference score between the incongruent and congruent trials. Although data analysis is still ongoing, preliminary results suggest that, compared with the unstandardized SNARC slopes in Study 6, outcome measures for symbolic and non-symbolic magnitude comparison on subsequent days appear to be more strongly related to one another. Nevertheless, overall, these other outcome variables also seem to vary substantially both between and within subjects, much like the unstandardized SNARC slopes, in terms of both their medians and their variability.

Conclusion. The SNARC effect reveals strong inter- and intraindividual variability, suggesting that it is an indominant and temporally unstable phenomenon. As claimed by Judd et al. (2023), “variability is a fundamental part of the human phenotype,” and researchers should therefore keep in mind both differences between individuals and fluctuations across time points within individuals when interpreting cognitive effects. When reporting on psychological research, it is important to explicitly state whether interpretations of collected data are drawn at the group level or the

individual level, and, at least in the latter case, whether anything is known about their temporal stability.

3.2. Implications for theoretical accounts of the SNARC effect

After integrating the results from all six studies, I will outline how the findings relate to theoretical accounts of the SNARC effect. Specifically, I will discuss whether the MNL account, the WM account, as well as the polarity and markedness accounts can explain the findings.

3.2.1. The Mental Number Line account

As outlined in the Introduction, the SNARC effect was first explained with a MNL in the study of its discovery (Dehaene et al., 1993). Many results from CJ, PJ, and MC presented in this dissertation can indeed be explained by an underlying MNL, although findings differed between tasks. Specifically, a SNARC effect was found in each task, but differences in strength and direction were observed. Assuming an underlying MNL, these findings suggest either that the MNL is more or less strongly activated depending on current task demands, or that several versions of the MNL with differing resolutions can be activated depending on the task. For instance, the MNL might be less strongly activated in CJ than in MC, or a MNL with lower resolution might be activated in CJ compared to MC. Notably, the MNL is not necessarily linear, meaning that each mathematical interval in terms of numerical distance is not always reflected by the same spatial interval in a human mind proportionally. Instead, the MNL may be logarithmic (Dehaene, 2001; Izard & Dehaene, 2008), with high precision for smaller numbers but compression for larger numbers. A non-linear MNL could explain that the SNARC effect in CJ (Study 1) was only driven by numbers 1, 2, and 3, and it also accounts for the descriptively steeper slope for small versus large numbers in PJ (Studies 2 and 3). To summarize, many empirical findings are in line with the MNL account. Yet, the MNL account (at least with its most common assumptions) cannot explain all observations made in the studies included in this dissertation, nor all findings from other studies. Below, I will therefore outline dissonances between typical claims of the MNL account and empirical findings, while suggesting additional requirements and extensions for the MNL account.

The MNL aligning with script direction. To start with, the direction of an individual's MNL is said to align with the script direction of their native language. From this claim, one would predict an association of small numbers with the left and large

numbers with the right in LR-language speakers, and the reversed association in RL-language speakers. However, results for the Iranian sample tested in Study 3 did not support this claim. We found a regular SNARC effect in PJ and could not detect any effect in MC with a descriptive tendency towards a regular SNARC effect. Several other recent articles reported similar results (e.g., Bulut et al., 2024; Hochman et al., 2024; Lopiccolo & Chang, 2021). If a MNL exists, its direction might not rely solely on script direction of the native language. The origins of the MNL must be extended to other factors, such as script direction in non-native languages that became relevant through globalization and technology, or other cultural directional preferences. Moreover, if the mental representation of numbers follows a MNL, its direction might rely on number-related directional conventions rather than language-script directions in written texts. That is, in all native RL languages of participants tested so far (i.e., Arabic, Farsi, Hebrew), numbers are written and read in a LR fashion (e.g., in three-digit numbers, the hundreds digit is the leftmost, the tens digit is in the middle, and the units digit is the rightmost). Similarly, in mathematical graphs, data plots, and simple coordinate systems introduced in school, number magnitude increases from left to right. Thus, the MNL account would be improved by deriving it not only from language-script direction, but also from number-related directional habits. Note, however, that even with these extensions, the MNL account might not be able to explain SNARC-like effects observed in newborn humans or non-human animals (Gazes et al., 2017; Rugani et al., 2015; Rugani & de Hevia, 2017). The latter are not influenced directly by written texts or language. At most, one could assume indirect influences, for example, observations of adult humans' reading and writing behavior.

The MNL being highly flexible. Another property of the suggested MNL is its high flexibility, supported by the finding that the same numbers can be associated with the left when they are the smallest in a stimulus set, but with the right when they are the largest (Dehaene et al., 1993; Fias et al., 1996). While we were able to replicate this RM when investigating PJ with different number ranges (Study 5), we also found a more pronounced SNARC effect in the lower range compared to the higher one, suggesting a simultaneous role of AM. We therefore argue that the SNARC effect depends on both RM and AM, which are not mutually exclusive. Hence, zooming in and out a linear MNL seems only partly plausible (Pinhas et al., 2013), but only to a certain extent. If the underlying MNL was assumed to be non-linear, however, zooming in and out might be fully possible.

The MNL underlying number processing specifically. While the MNL, as its name suggests, refers only to the mental representation of numbers, it should also be extended to other ordinal stimuli. Otherwise, the MNL account cannot be applied to paradigms such as the CVC task using letter stimuli (Study 4). Numbers may represent only one instance among many spatially represented kinds of magnitude or ordinality. Interestingly, letter stimuli reveal a similar pattern to that described for numbers above: The letter A, and potentially subsequent letters with early ordinal positions in the alphabet, is most strongly associated with the left side. This may point to a logarithmic shape, as is the case for numbers.

Conclusion. Despite the MNL account's continued popularity, several theoretical extensions would be necessary for it to explain all empirical findings.

3.2.2. The Working Memory account

As outlined in the Introduction, a controversy exists regarding whether WM or LTM content determines the mental representation of numbers and other ordinal stimuli. As will be discussed below, we found evidence for both sides.

Ordinality within the current number context in WM. Study 5 provides interesting insights using numbers as stimuli. On the one hand, the number range currently used as the stimulus set clearly influenced spatial response-side preferences. This was the case even though participants were neither informed about the stimuli they would be presented with nor instructed to memorize them. The association of the same numbers (i.e., 4 and 5) with the right when they were the largest numbers in the range, and with the left when they were the smallest, suggests that RM determined behavior. On the other hand, the SNARC effect was stronger for small numbers compared to large ones (see the conceptual replication in Experiment 2), and there was a disproportionally strong left-hand advantage for the number 1 in the lower range. This indicates that both RM and AM contributed to the mental representation of numbers. Crucially, RM can be interpreted as ordinality stored in WM, whereas AM can be interpreted as magnitude or ordinality stored in LTM. In other words, the debate about a flexible, RM-dependent versus a fixed, AM-dependent SNARC effect can be understood as part of the broader debate about a WM-based versus LTM-based SNARC effect. To summarize, Study 5 provides mixed evidence regarding the WM account. The originally proposed WM account fails to explain all our findings, as doing so requires an account that incorporates both a flexible and a fixed component. Only the revised version of the WM account (Abrahamse et al., 2016), which proposes that

a spatial template in WM is filled in during the task, with ordinality stemming from either WM or LTM, or a hybrid account (van Dijck et al., 2015) can explain our findings.

Alphabetical ordinality of letters in LTM. Apart from using different number ranges in Study 5, using different kinds of stimuli can also shed light on the controversy, one which is closely related to the debate about the role of magnitude and ordinality. With numbers, disentangling magnitude from ordinality is hard (although possible, as done by Koch et al., 2023), because the two usually coincide. In contrast, tasks using letters offer a different perspective, because letters entail ordinality but not magnitude. We therefore used letters as stimuli in Study 4, carefully manipulating whether ordinality stemmed from LTM (i.e., alphabetical order) or was newly acquired in WM (i.e., an arbitrary memorized sequence). Unfortunately, past research often failed to disentangle or explicitly describe whether ordinality came from LTM or WM, and how salient each source was made during the task, which we sought to enhance in Study 4. We observed an OPE based on LTM in all three experiments, consistent with previous studies (Gevers et al., 2003, 2004). Notably, no such effect was detected in the seminal SNARC study (Dehaene et al., 1993), which led the authors to conclude that magnitude instead of ordinality was decisive for the SNARC effect in numbers. However, that study may simply have missed a true underlying effect due to its very small sample size and lacking statistical power. Moreover, Study 4 revealed stimulus-specific effects like an especially strong association of letter A with the left across all three experiments. These findings are not in line with the WM account in its original form, but they can be reconciled with the WM account in its revised form, which proposes that a spatial template in WM is filled in with ordinal stimuli during task execution, while ordinality can be retrieved from WM or LTM.

Newly memorized ordinality of letters in WM. In addition to the OPE based on the alphabetical order found in all three experiments in Study 4, we also detected an OPE based on a newly learned letter sequence, which aligns with the WM account. However, the WM-based OPE was only elicited in Experiment 3, which required regular retrieval of the letter sequence in WM, but not in Experiment 2, which merely required its maintenance. In line with a similar study by Ginsburg et al. (2014) using numbers, this finding suggests that the memorized sequence needs to be regularly activated in order to affect behavior. That is, high WM salience is required to elicit spatial response preferences based on WM, and rather than overruling spatial response preferences based on LTM content, both effects can coexist. The WM account does not include the

salience of WM content as a prerequisite for observing a WM-based OPE and would therefore need to be extended to account for these observations.

Further evidence for WM involvement. The WM account is also underpinned by the finding that different parts of WM seem to be involved in typical SNARC tasks in any case. While the SNARC effect in MC appears to rely more strongly on visuospatial WM, it seems to rely more on verbal WM in PJ (Deng et al., 2017; Herrera et al., 2008; van Dijck et al., 2009). Moreover, the SNARC effect is prone to situational influences (Cipora et al., 2018), such as the imagination of a number context (e.g., a ruler with linearly increasing magnitudes starting from the left vs. a clock with circularly increasing magnitudes starting from the right; Bächtold et al., 1998, but see Mingolo et al., 2023), or priming by different spatial positions of numbers in a text (Fischer et al., 2010). Such influences on the SNARC effect can only be explained through WM involvement, not by sole reliance on LTM content.

Conclusion. Our results from Studies 4 and 5 cannot be fully explained by the WM account alone. Instead, in line with previously published literature (e.g., Ginsburg et al., 2014; Koch et al., 2023), they support a notion of mixed WM and LTM influences on spatial mental representations of magnitude and ordinality.

3.2.3. Polarity and markedness accounts

The Polarity Correspondence principle (Proctor & Cho, 2006), the Linguistic Markedness account for numerical cognition (Nuerk et al., 2004), and the Psychological Markedness account (Nuerk & Schroeder, 2024) all claim that the SNARC effect is a compatibility effect between the polarity or markedness status of stimuli and response sides. For example, assuming that the stimulus feature “small” entails negative polarity and is linguistically or psychologically marked, and assuming the same for the response side “left,” a trial where a left-handed response must be given to a stimulus number 1 would reflect a compatible scenario. In the following, I will outline that the Psychological Markedness account does better in explaining our findings than the Polarity Correspondence principle and the Linguistic Markedness account do.

Categorical versus graded features leading to different SNARC effects. In contrast to psychological markedness, linguistic markedness and polarity were originally framed as categorical rather than continuous (Proctor & Cho, 2006). The categorical shape of the SNARC effect that we had predicted in MC aligns with this framing: All numbers that fall into the “small” category were expected to be associated

with the left, and all numbers that fall into the “large” category with the right. However, the RT patterns we observed in Studies 2 and 3 revealed a mixture of a categorical and a continuous shape, which does not align with the dichotomous concepts postulated in the Polarity Correspondence principle and the Linguistic Markedness account. Notably, even the Psychological Markedness account, which conceives of markedness as graded and task-dependent, may fail to explain this mixture of a categorical and a continuous shape in MC: Whenever participants are instructed to dichotomize the stimulus set as according to task instructions, psychological markedness may also be dichotomous. As concerns the SNARC effect observed in other tasks, continuous or logarithmic shapes are not in line with purely categorical claims such as the Polarity Correspondence principle and Linguistic Markedness accounts. However, they may not contradict the Psychological Markedness account. For instance, although the difference between 1 and 2 is mathematically equal to the difference between 8 and 9, the former may be perceived as larger than the latter in some tasks, as accounted for by the context-dependence of psychological markedness. To conclude, non-categorical shapes of the SNARC effect as found in PJ (Studies 2, 3, 5, and 6) or CJ (Study 1) are not in line with the Polarity Correspondence principle or Linguistic Markedness account, and the Psychological Markedness account also struggles with explaining them at least in MC.

Flexibility of polarity and markedness. Importantly, as explained earlier, the mixed findings from Studies 4 and 5 require theories to include an explanation for fixed, LTM-based as well as for flexible, WM-based parts of the SNARC and SNARC-like effects. To my knowledge, the concepts of polarity and linguistic markedness do not incorporate the latter, and only psychological markedness is claimed to adjust to the context. The logic is as follows: For the example of Experiment 2 in Study 5 using different number ranges for PJ, when numbers from 1 to 5 are used, numbers 4 and 5 might be psychologically unmarked, whereas they might be marked when numbers from 4 to 8 are used. Apart from that kind of flexibility, we observed high inter- and intraindividual variability for the SNARC, MARC, and Odd effects in Study 6 (see also Cipora et al., 2019), which should be incorporated in theories aiming at explaining these effects. Only a minority of the participants reveals the effects, some participants even show reversed effects, and in fluctuations between sessions within the same participants are high. The Psychological Markedness account considers individual differences and contextual influences on cognitive effects, while the Linguistic

Markedness account or Polarity Correspondence principle claim more general principles.

Polarity or markedness leading to other associations. Beyond the SNARC and SNARC-like effects, polarity and markedness also play a role for other cognitive effects. That is, stimulus features apart from magnitude (small vs. large) or ordinality (early vs. late) leading to the SNARC and SNARC-like effects can also entail polarity or markedness and therefore lead to other compatibility effects with response sides. Below, I will explain how three example stimulus features other than magnitude or ordinality can coincide or not with response sides in terms of their polarity or markedness status: the parity status of numbers (odd vs. even), the luminance level of different kinds of stimuli (light vs. dark), or the nature of letters (consonant vs. vowel).

First, as explained earlier, the polarity and markedness accounts can explain the categorical MARC effect based on the compatibility between “odd” and “left,” being marked and entailing negative polarity, and between “even” and “right,” being unmarked and entailing positive polarity. To our surprise, we could not detect the MARC effect in Study 2 in PJ despite task instructions explicitly referring to parity. At the same time, we found little evidence for the MARC effect in Study 1 in CJ where parity was task-irrelevant. Note that we tested participants with LR native languages in both studies. The null findings obtained in Study 2 were unlikely due to low statistical power for small effect sizes or methodological flaws, because we tested a large sample in the most common experimental PJ setup. The Polarity Correspondence principle and the Linguistic Markedness account fail to explain the lacking MARC effect in Study 2. According to the Psychological Markedness account, the null findings could in theory be due to individual differences (e.g., many participants happen to be part of the tested sample who do not have strong mental representations of parity). However, I consider this explanation as unlikely. Nevertheless, in the cross-cultural investigation of PJ in Study 3, we found a MARC effect in German participants, but not in Turkish or Iranian participants. This is in line with formal linguistic markedness: In German, the marked word “odd” (“ungerade”) is a direct negation of the unmarked word “even” (“gerade”), whereas no such obvious linguistic markedness might be present in the Turkish (“tek” and “çift”) and Farsi (“fard” and “zoj”) words. To conclude, the presence of the MARC effect in Studies 1 and 3 align with the three polarity or markedness accounts, whereas its absence in Study 2 does not.

Second, we found an association of light with the left and dark with the right, in line with (Ren et al., 2011), but not with Fumarola et al. (2014), Gusmão et al. (2022), Löffler et al. (2022), and Wang et al. (2022). This finding suggests that light is marked, and dark is unmarked, although the direction of this association contradicts the direction of previously reported associations. The Polarity Correspondence principle and the Linguistic Markedness account fail to account for the opposing patterns, whereas the Psychological Markedness would be able to explain them if we knew of contextual factors that differed between experimental setups in the mentioned studies.

Third, the nature of letters in terms of whether they are consonants or vowels might have a polarity or markedness status as well. Moreover, the height of letters is claimed to cause phonological markedness (Rice, 2007): High and low vowels seem to be unmarked, whereas mid vowels seem to be marked. Regarding psychological markedness, differences may also exist between individuals based on individual experience with specific letters, such the initials in first and last names. Note that we did not check compatibility effects between polarity and markedness of any of these letter features with response sides in Study 4. It would certainly be interesting for future research to investigate them, although larger stimulus sets would be required.

Conclusion. In sum, the dichotomization of stimulus features postulated in the Polarity Correspondence principle and the Linguistic Markedness account is not in line with most of our findings. Both accounts fail to explain the partly flexible nature of the SNARC and SNARC-like effects revealed when using different stimulus sets, as well as their inter- and intraindividual variability. While the Psychological Markedness account postulates more nuanced and flexible cognitive mechanisms and therefore seems more compelling at first glance, it does not make precise and correct predictions for the SNARC and SNARC-like effects in all experimental setups either.

3.2.4. Other theoretical accounts

Overall, findings from all six studies demonstrate that the SNARC effect and the OPE are complex phenomena. Their strength and shape depend on task instructions (e.g., task relevance of magnitude or ordinality; Studies 1, 2, and 3), and sample characteristics (e.g., native language and culture; Study 3). Moreover, the SNARC effect and the OPE appear partly flexible, though not entirely: to some extent, they adjust to stimulus sets (e.g., number ranges; Study 5) and to secondary task demands (e.g., retrieval from WM; Study 4). Simplistic accounts therefore struggle to explain the SNARC effect and the OPE across all contexts in which they are observed. Instead,

more nuanced theories that address multiple influences, such as the hybrid account (van Dijck et al., 2015) and the multiple-coding framework (Schroeder et al., 2017c), offer more plausible explanations. Note, however, that even those theories have hardly addressed the pronounced interindividual variability of the effects (Cipora, van Dijck, et al., 2019), and even less their potentially high intraindividual variability (as observed for a small sample in PJ in Study 6). Findings concerning their prevalence and temporal stability highlight that researchers should refrain from the default assumption that cognitive effects are dominant phenomena (Rouder & Haaf, 2018) or that they reflect trait-like representations. Instead, researchers should consider that cognitive effects may just as well be indominant phenomena or reflect state-like mental representations.

Having discussed insights from various tasks, cultures, stimuli, and time points with respect to theoretical accounts of the SNARC effect, it is crucial to note that additional cognitive effects can be elicited within the presented paradigms. Indeed, several studies within this dissertation revealed other typical numerical-cognition effects that can interact with the SNARC effect. For instance, Study 2 demonstrated a Numerical Distance Effect (slower responses for larger compared to smaller distances between the stimulus number and the reference number; Moyer & Landauer, 1967): Overall, discriminating 4 and 6 from 5 took longer than discriminating 1 and 9 from 5. Furthermore, the SNARC effect is typically stronger in slower than in faster responses (Didino et al., 2019). Together, these two effects may reduce spatial biases for more extreme numbers in a stimulus set (e.g., 1 and 9), while amplifying them for more centered numbers (e.g., 4 and 6), thus making the SNARC effect's shape appear more categorical (Gevers, Verguts, et al., 2006).

To conclude, the SNARC effect and the OPE are complex phenomena, which require theoretical accounts to consider a wide range of influencing factors. Moreover, behavioral observations may result from their interaction with other cognitive effects.

3.3. A computational contextualization model for the SNARC and SNARC-like effects

The empirical findings presented in this dissertation provide new insights into processing of magnitude or ordinality and their spatial mental representations. Several important aspects, like their high automaticity, especially for small numbers (Study 1), different shapes and strengths depending on task instructions (Study 2), and the roles of WM and LTM therein (Study 4) were shown. Although many theoretical accounts have been proposed for the SNARC and SNARC-like effects, they cannot explain all

empirical findings or stay vague about underlying cognitive mechanisms. As outlined in the Introduction, computational modeling is a helpful approach for connecting behavioral observations with assumptions on underlying cognitive factors. Butz et al. (2025) relate this approach to the idea that contextualization is at the core of cognition. The authors claim that contextualization is necessary for the human mind to efficiently plan and control actions depending on what is task-relevant. On the one hand, contextualization refers to the intentional encoding of task-relevant features of the environment. For instance, in an experimental testing environment designed for assessing the SNARC or SNARC-like effects, task-relevant features such as font color or parity status of the presented number stimuli are defined in the task instructions. On the other hand, contextualization also includes the automatic activation of context schemata, which refer to cognitive situational templates. For example, during investigations of the SNARC or SNARC-like effects, magnitude or ordinality schemata may be activated. Cognitive context schemata include predictive biases, which may reflect associations between the presented stimuli and horizontal space in this example, such as a strong left association for stimuli like number 1 or letter A. With the idea in mind that the simultaneous activation of different cognitive context schemata may elicit the SNARC and SNARC-like effects, we are currently developing a computational model, which I will introduce in the following. We implement the model in the Python language (Python Software Foundation, 2025; for details, see Appendix A7.1.) with the help of Claude AI (Anthropic, 2025)[6].

3.3.1. Aims and basic model architecture

The aim of this project is to create a nuanced model that accounts for the SNARC and SNARC-like effects in a variety of experimental settings. While other accounts only explain basic SNARC and SNARC-like effects within the most common experimental settings and remain vague regarding parameter specifications, we are designing our computational model to be more fine-grained and explicit. Seeing the SNARC and SNARC-like effects through the lens of contextualization, we are testing for the potential influence of several cognitive context schemata on behavioral patterns

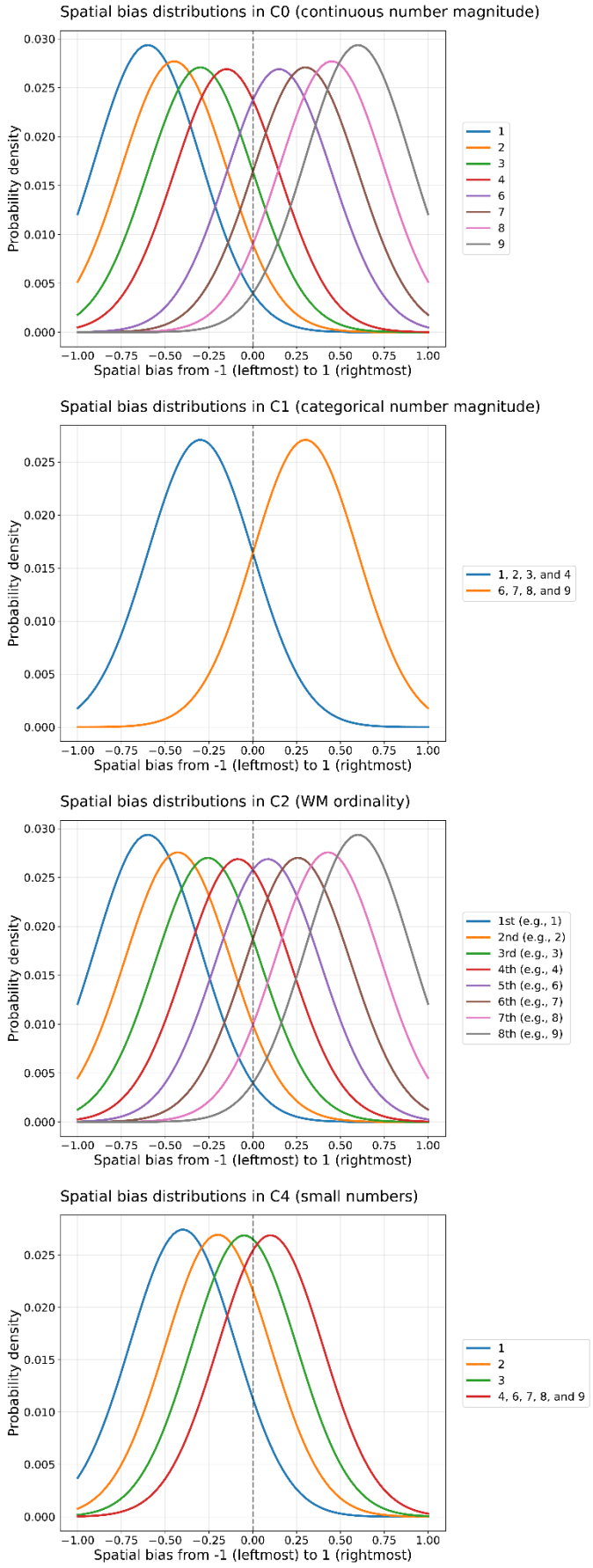
[6] This project is a collaboration with M. V. Butz, K. Cipora, G. Hassan, and H.-C. Nuerk. We used Claude AI (Anthropic, 2025) to support the technical implementation of the computational model in Python (Python Software Foundation, 2025). The initial scaffolding of the code was generated with Claude AI, after which we used it primarily for structuring, debugging, and exploring different algorithmic approaches. All theoretical foundations and conceptual decisions about the model and its framework were made independently. The data preparation, the interpretation of results, and the writeup for this dissertation were carried out without AI assistance.

in different experimental tasks. The current version of our computational model works for six tasks, three of which use number stimuli from 1 to 9 (excluding 5), and three of which use letter stimuli A, E, J, L, O, R, U, and W. Specifically, the number tasks we have implemented so far are nominal CJ (as in Study 1, Experiment 1), MC (as in Studies 2 and 3), PJ (as in Studies 2, 3, and 6), and the letter tasks are CVC with low, intermediate, and high WM salience (as in Study 4, Experiments 1, 2, and 3).

Crucially, the model builds upon the assumption that a combination of cognitive context schemata is activated in each task, eliciting the SNARC and SNARC-like effects. The five cognitive context schemata that we include in the current version are continuous magnitude for numbers (C0), categorical magnitude for numbers (C1), WM ordinality for any stimulus kind (C2), LTM ordinality in the alphabet for letters (C3), and a bias for small numbers (C4). While these five schemata, illustrated in Figures 3-2 and 3-3, appear plausible to us, alternative ones can be considered. We want to assess whether such schemata are generally able to account for the most common SNARC and SNARC-like effects. For each cognitive context schema, we defined a set of predictive biases (Butz et al., 2025), namely cognitive spatial biases in the form of probability density distributions (henceforth “spatial bias distributions”), ranging from −1 (strongest possible bias towards the left) to 1 (strongest possible bias towards the right). These reflect our theoretical assumptions on associations of specific stimuli with horizontal space, which may cause facilitation or inhibition of left- or right-sided responses. While not directly measurable, we selected the bias distributions based on our empirical knowledge about the SNARC and SNARC-like effects. We expect them to be activated differently across tasks, causing different strengths and shapes of the SNARC and SNARC-like effects. The spatial bias distributions in C0, C1, C2, and C4 for number tasks are illustrated in Figure 3-2, and those in C2 and C3 for letter tasks are illustrated in Figure 3-3. A detailed description of these spatial bias distributions, as well as specifications regarding their technical implementation can be found in Appendix A7.1.

Figure 3-2

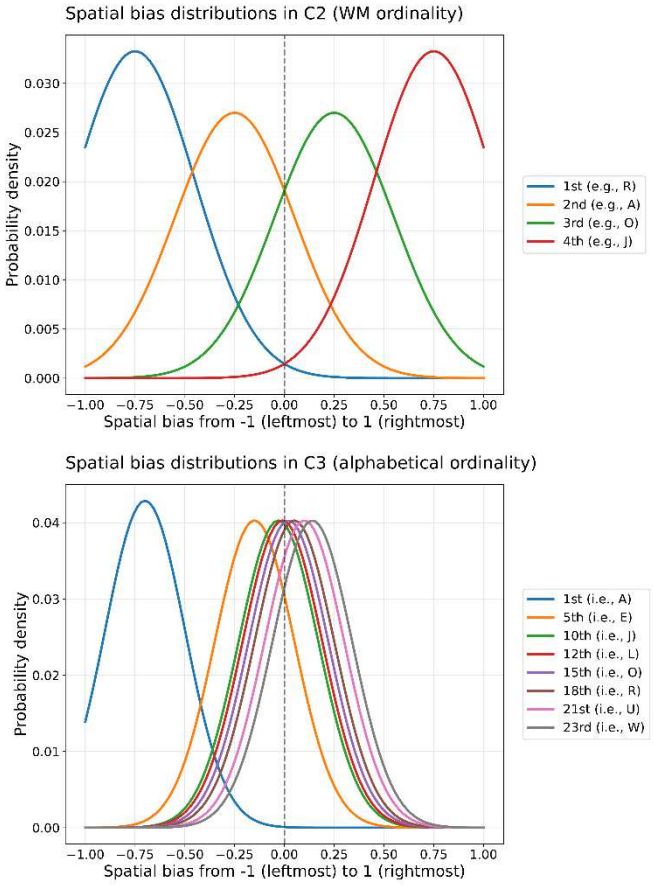
Spatial bias distributions in contexts that may be activated in number tasks



Note. This figure was created in Python (Python Software Foundation, 2025). It illustrates the spatial bias distributions in contexts C0, C1, C2, and C4, which can be activated to varying extents in tasks using number stimuli (i.e., CJ, MC, and PJ). The probability density is plotted as a function of the spatial bias, with −1 reflecting the maximal left bias and 1 reflecting the maximal right bias. By definition, the area under each probability distribution equals 1, and the truncation at −1 and 1 causes the maximal probability density to differ between distributions.

Figure 3-3

Spatial bias distributions in contexts that may be activated in letter tasks



Note. This figure was created in Python (Python Software Foundation, 2025). It illustrates the spatial bias distributions in contexts C2 and C3, which can be activated to varying extents in tasks using letter stimuli (i.e., CVC with low, intermediate, and high WM salience). The probability density is plotted as a function of the spatial bias, with −1 reflecting the maximal left bias and 1 reflecting the maximal right bias. By definition, the area under each probability distribution equals 1, and the truncation at −1 and 1 causes the maximal probability density to differ between distributions. Note that distributions in C2 are plotted for one of the stimulus sets we used in Study 4, and that WM ordinality can either be based on the sequence stored in WM (e.g., R, A, O, J), or, if no sequence is stored in WM, on the alphabetical order stored in LTM ranked within the current set (e.g., A, J, O, R).

The most crucial functionality of the computational contextualization model is that it simulates trial-level datasets for up to six different tasks, user-defined sample sizes, and numbers of stimulus repetitions per experimental cell. It additionally preprocesses this trial-level data, analyzes and plots sample-level SNARC and SNARC-like effects, and stores the results in external files. For details on functionalities of our Python code and the data simulation mechanism that we designed and implemented so far, please refer to Appendix A7.2. Notably, by slightly editing the code, cognitive context schemata can be added or removed, and spatial bias distributions within the existing ones can be modified as the project progresses.

To determine the extent to which the proposed cognitive context schemata encoding the spatial biases may contribute to the SNARC and SNARC-like effects, we optimize their relative weights in each of the six tasks (for default weights, see Table A7-1 in Appendix A7.3.). In other words, we aim at fine-tuning the impact of each cognitive context schema in each task, such that the simulated data predicted by the model matches empirical data. The empirical data we supplied the model with as of now was the data we collected in Studies 1 (nominal CJ), 2 (MC and PJ), and 4 (CVC with low, intermediate, or high WM salience) of this dissertation. To account for large interindividual variability of the SNARC and SNARC-like effects (as found for MC and PJ; Cipora et al., 2019; Hohol et al., 2020, Supplementary Materials), we ran the optimization at the individual level, separately for each participant. In the following, I will focus on optimized weights for the proposed cognitive context schemata. Note, however, that we also optimized further parameters for each participant: their baseline and standard deviation (SD) of RT, and their sensitivity to cognitive conflict, reflected in how strongly the combined spatial bias interferes with the correct response side (for details and an alternative approach, see Appendix A7.3.). The optimized weights for the proposed cognitive context schemata help to better understand their roles (i.e., by reflecting their level of activation) in each task. We expect them to provide crucial insights into cognitive mechanisms that underlie the SNARC and SNARC-like effects.

3.3.2. Preliminary optimization results

In this section, I will report on the most crucial results obtained in the parameter optimization (for additional results from this and an alternative approach, see Appendix A7.4.). Note that this project is still in progress, and all reported results should therefore be considered as preliminary. Overall, the parameter optimization led to a substantial improvement in each of the six tasks (for details, see Table A7-2 in Appendix A7.4.).

This indicates that the computational contextualization model can predict behavioral patterns better when using optimized parameters compared to default ones. The main results, namely the optimized weights for the cognitive context schemata that we postulated, are displayed in Table 4 (for more optimized parameters, see Tables A7-3 and A7-4 in Appendix A7.4.). It is important to note that the optimization algorithm we currently use yields extreme boundary values (i.e., all context weights were 0 except one, which was 1) for many participants. As it seems implausible that only one cognitive context schema at the time is activated in a participant in a certain task, I will focus on averaged parameter estimates across participants when interpreting the preliminary results in this dissertation. Importantly, we will explore different optimization algorithms in the future and expect to obtain context weights that reflect multiple schemata being activated simultaneously.

Starting with the number tasks, all available schemata seem to play an equally important role in nominal CJ. According to our results in Study 1, one might have expected C4 to be most decisive, but this hypothesis was not supported by these preliminary optimization results. The picture looks different for MC and PJ. Surprisingly, C0 (i.e., continuous number magnitude) and C2 (i.e., WM ordinality) do not seem to play a big role in either of both tasks requiring semantic number processing. Instead, more than half of the weight lies on C1 (i.e., categorical number magnitude) in MC, which is in line with the categorical shape described in the literature (Wood et al., 2008) and revealed in Study 2. C1 also seems to partially predict response patterns in PJ. However, the schema with the strongest weight in PJ turned out to be C4 (i.e., left biases for small numbers). This finding is interesting because most theoretical accounts assume continuous mental number representations (e.g., the MNL account, the WM account, and the Psychological Markedness account), and some assume categorical ones (e.g., the Polarity Correspondence principle, and the Linguistic Markedness account). We therefore expected a stronger influence of C0 or C2 (i.e., continuously spread spatial bias distributions) or of C1 (i.e., dichotomization of stimuli in spatial bias distributions) than of C4. However, the result is plausible because the optimization was run on data from Study 2, where the SNARC slope indeed appeared steeper in the lower than higher range of the stimulus set (see Figure 5, Panel B, in Study 2).

Table 3-1

Optimized weights for all schemata in the computational contextualization model

	Mean weight (SD)				
Task	**C0**	**C1**	**C2**	**C3**	**C4**
Nominal CJ	0.240 (0.123)	0.273 (0.129)	0.245 (0.120)	0 (0)	0.241 (0.126)
MC	0.042 (0.123)	0.573 (0.455)	0.125 (0.288)	0 (0)	0.260 (0.413)
PJ	0.026 (0.087)	0.216 (0.367)	0.070 (0.194)	0 (0)	0.688 (0.412)
CVC (low)	0 (0)	0 (0)	0.247 (0.410)	0.753 (0.410)	0 (0)
CVC (interm.)	0 (0)	0 (0)	0.836 (0.327)	0.164 (0.327)	0 (0)
CVC (high)	0 (0)	0 (0)	0.868 (0.310)	0.132 (0.310)	0 (0)

Note. For the purpose of summarizing general tendencies, this table displays the means and SD for the weights across participants, although they were optimized at the individual level. These weights refer to the relative influence of the respective cognitive context schema onto the response pattern in each task. The weights were fixed at 0 for C3 in number tasks and for C0, C1, and C4 in letter tasks. The CVC descriptions in parentheses in the leftmost column refer to low, intermediate, and high WM salience. An extended version of these results can be found in Table A7-3 in Appendix A7.4.

For the three letter tasks, C2 (i.e., WM ordinality) and C3 (i.e., LTM ordinality in the alphabet) are currently the only two cognitive context schemata that we consider in the computational contextualization model. Depending on task instructions in CVC, spatial bias distributions in C2 refer to different kinds of WM ordinality. Specifically, when no letter sequence needs to be memorized (i.e., low WM salience), the biases refer to the ordinal rank in the alphabet relative to other letters contained in the stimulus set (e.g., if A, J, O, R is the stimulus set: A = 1st, J = 2nd, O = 3rd, R = 4th). In contrast, when a letter sequence needs to be memorized (i.e., intermediate or high WM salience), the biases reflect the ordinal positions within that sequence (e.g., if RAOJ is the sequence: R = 1st, A = 2nd, O = 3rd, J = 4th). In line with our expectations, the postulated alphabetical context relying on LTM seems more strongly activated compared to the postulated WM ordinality context when WM salience is low, and the

opposite is true when WM salience is intermediate or high. However, these preliminary results for letter tasks may partially represent an optimization artifact in addition to truly underlying cognitive mechanisms. Specifically, the randomized assignment of participants to go stimulus sets (in CVC with any degree of WM salience) and memorized sequences (in CVC with intermediate and high WM salience) in the simulation procedure creates mismatches between the real and simulated experimental setups. For instance, data for a participant who learned the sequence RAOJ in the real experiment may be simulated as having learned LEWU instead. Nevertheless, given that only two go stimulus sets and four letter sequences were used in total, a proportion of the data was simulated for the truly used experimental setup. Moreover, the clear difference between the context weights in CVC with low WM salience and those in CVC with intermediate or high WM salience are in line with our findings in Study 4. Consequently, these results may not merely reflect an artifact. We will perform more checks in the future and further improve the implementation of the optimization.

Great interindividual differences were found in optimized weights of cognitive context schemata, as reflected by large SD for the means (see Table 3-1). While these differences may reflect an inadequate optimization procedure yielding boundary values for many participants, they may additionally be caused by actual high interindividual variability underlying the SNARC and SNARC-like effects (Cipora et al., 2019; Hohol et al., 2020, Supplementary Materials). Moreover, high intraindividual variability of the effects may contribute to the large SD we obtained, because combinations of cognitive context schemata determining response patterns may even differ between time points within participants (see Study 6). In further steps of this project, we may therefore also optimize model parameters using longitudinal data. Hence, minimizing the SD of the mean weights may not be a meaningful objective in this project.

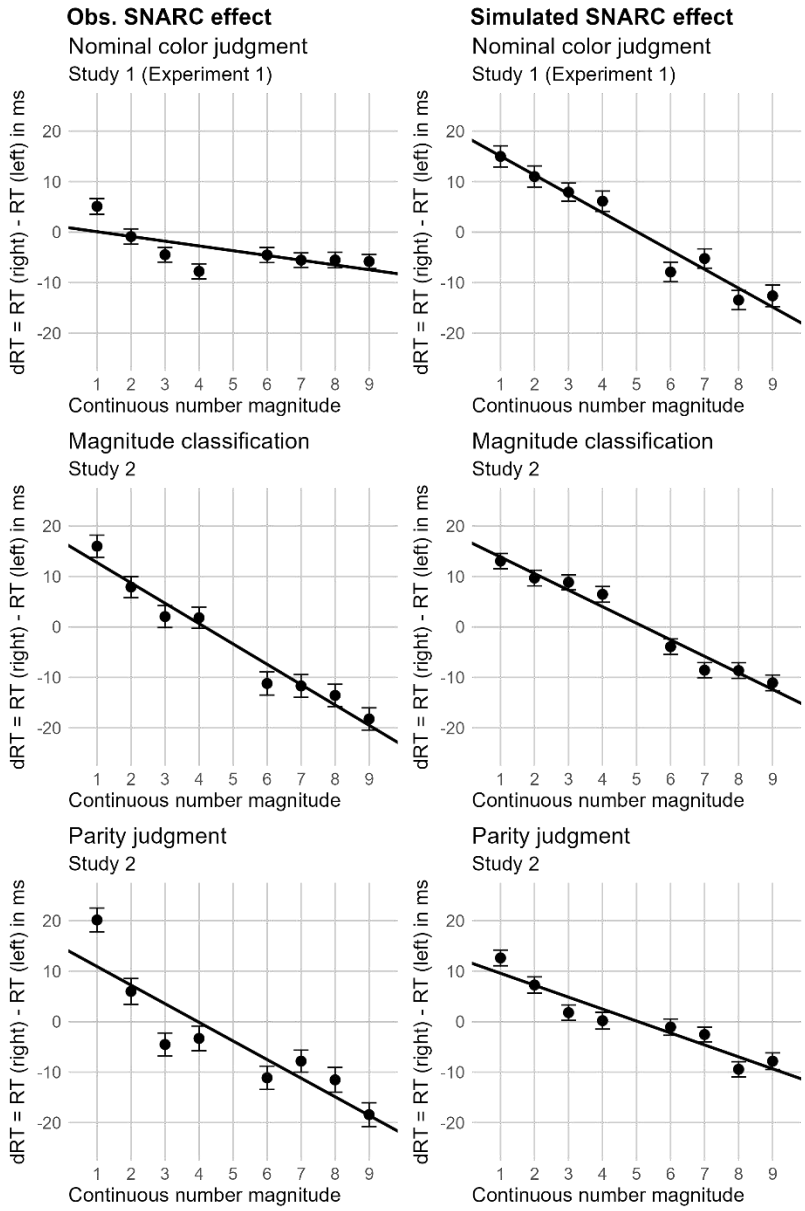
A robustness test was run in order to assess whether the optimization produces stable results across runs, results of which must be considered when drawing conclusions about model outputs such as the high interindividual variability. For details and exact results, please refer to Tables A7-5 and A7-6 in Appendix A7.4. Overall, optimized context weights were relatively stable, as suggested by similar estimates for “digital twins” (i.e., data generated multiple times under the same conditions). However, weights varied more in CJ than in other tasks, and outcomes may be hardly interpretable in CVC due to the mismatch of letter sets and sequences between the

simulation and the real experimental setup. The objective function also produced stable results, indicating similar quality of the optimized parameters across runs.

Finally, we simulated data from the computational contextualization model for each task based on the preliminary optimization results. The dRT patterns aggregated across participants within each number task are plotted in Figure 3-4, so that the group-level SNARC effects between observed and simulated data can be compared by the reader. For similar visualizations and comparisons of dRT patterns in letter tasks, and for further ones using a different parameter optimization approach, please refer to Figures A7-2, A7-3, and A7-4 in Appendix A7.5. The plots demonstrate that the simulated data clearly reveals a SNARC effects in each number task. Additionally, more subtle patterns in the empirically observed data are identifiable in the simulated data: For example, the SNARC slope appears steeper in the lower half of the stimulus set in each task, and a tendency towards a categorical SNARC effect is visible in MC. However, other characteristic behavioral patterns were not yet captured by our model, such as particularly large dRT for number 1 in each task as well as the overall weaker SNARC effect in CJ. We are confident that we will be able to make simulated data even more similar to truly observed data when further improving the optimization procedure as this project progresses.

Figure 3-4

Comparison of observed and simulated SNARC effects in three number tasks



Note. This figure was created in R (R Core Team, 2025). It shows six plots, each of which is a typical illustration of the SNARC effect in tasks using numbers as stimuli. Therein, dRT averaged across all participants (with bars reflecting standard errors of the mean dRT) on the y-axis are shown as a linear function of continuous number magnitude on the x-axis. Plots in the left column show the empirically observed data, whereas plots in the right column show the data simulated by the computational contextualization model. Plots in the upper, middle, and lower row illustrate data from nominal CJ, MC, and PJ, respectively. Note that these plots are also included in Figure A7-2, and similar plots for the LTM-based OPE and the WM-based OPE in letters are shown in Figures A7-3 and Figure A7-4 in Appendix A7.5.

3.3.3. Future model improvements and extensions

As shown in the previous section, the newly introduced computational contextualization model is able to simulate realistic RT data for six different tasks. The simulated data was obtained after first optimization attempts for weights of five postulated cognitive context schemata, and it reveals clear SNARC and SNARC-like effects. The simple architecture of the model, merely consisting in additive factors, has shown to be a promising approach. However, while data simulation and other basic functionalities function well already, the optimization algorithm must be further enhanced. Planned improvements and ideas for model extensions are outlined below.

As previously explained, it is important to identify the cause of the frequent boundary values for context weights at the individual level, as it seems implausible that they reflect truly underlying cognitive patterns. We may change the optimization settings or algorithm to prevent such implausible patterns. Moreover, we will modify the implementation of the data simulation, so that it assigns the truly run stimulus set and memorized sequence to each simulated participant in the letter tasks. While the randomly assigned stimulus set and memorized sequence do not pose any problem in the simulation itself, they preclude meaningful optimization for most participants.

Moreover, all reported optimization and simulation results are based on five cognitive context schemata. Although they reflect theoretical assumptions we made by carefully considering the literature on the SNARC and SNARC-like effects and our observations in several studies, alternative assumptions may be plausible. The schemata including the spatial bias distributions were tested for the very first time in the present investigation and may be modified or extended. Notably, some cognitive context schemata that currently hold only for numbers or letters may be combined to schemata that hold for several stimulus kinds. For instance, C3 for alphabetical ordinality in letters and C4 for small numbers may be combined to a schema with stronger, more spaced left biases for early items in a sequence and weaker, less spaced right biases for late items (i.e., potentially following a logarithmic pattern). Moreover, many more parameters can be optimized, such as the means and SD for spatial bias distributions, or accuracy rates in given responses. If adding these parameters leads to identifiability issues, they may at least be optimized across participants. We may also explore alternative optimizing algorithms and objective functions and the degree to which restricting parameter ranges further increases the robustness of optimization results.

The current version of the computational model already accounts for six tasks, including the most common number stimulus set as well as two exemplary letter stimulus sets. In the future, even more tasks (e.g., orientation judgment) and stimulus sets (e.g., different number ranges, see Study 5, or different stimulus kinds beyond numbers and letters entailing magnitude or ordinality) should be implemented. Moreover, the functionality of the Python code should be further improved for the user's convenience (e.g., by adding the option to compute a categorial instead of linear regression model for the analysis, see Study 2).

Apart from the planned model improvements and extensions described above, additional tests may be performed in the future to evaluate the validity of theoretical claims regarding the SNARC and SNARC-like effects. For instance, the model allows for entirely switching off certain context schemata (i.e., by fixing their weights to 0 during the optimization procedure). Most intuitively, to disentangle magnitude (i.e., LTM content) from ordinality (i.e., WM content) in number tasks, one may rule out the contribution of either C0 or C2 to behavioral patterns. By optimizing the weight of only one of these schemata and comparing the resulting optimization goodness with respect to the objective function, it becomes possible to investigate which of both schemata dominates in different tasks.

3.3.4. Embedding the model within existing theoretical accounts

To briefly recap, the newly introduced computational contextualization model proposes that the SNARC and SNARC-like effects result from the simultaneous activation of several cognitive context schemata (Butz et al., 2025). In this section, I will relate the schemata postulated within the model to existing theoretical accounts for the SNARC and SNARC-like effects outlined in Sections 1.2. and 3.2.

First, the model incorporates a cognitive context schema for continuous number magnitude (C0), the semantic meaning of which is stored in LTM. Similar to the most prominent version of the MNL account (Dehaene et al., 1993), namely the linear one, it postulates that spatial biases in number processing align with magnitude in a continuous manner, and that such biases in cognitive processing are specific to numbers. Accordingly, the model deactivates these biases for tasks using other stimulus kinds, such as letters.

Additionally, the model includes a schema for categorical number magnitude (C1), which dichotomizes the stimulus set, thereby resembling claims made by the Polarity Correspondence principle (Proctor & Cho, 2006), the Linguistic Markedness

account (Nuerk et al., 2004), or the verbal-spatial perspective (Gevers et al., 2010). While the current implementation is limited to numbers, future versions of this schema may extend to other stimulus kinds. That is, instead of referring specifically to numbers in terms of small and large, the schema could hold for many kinds of opposite stimulus features that may entail negative vs. positive polarity, or that may be marked vs. unmarked. Furthermore, adding graded markedness schemata, as discussed in the Psychological Markedness account (Nuerk & Schroeder, 2024), may be considered. However, this would only be meaningful if such additional schemata entail substantially different cognitive spatial biases from those already included in other schemata. For instance, if a newly postulated schema for psychological markedness of number magnitude included continuously spread spatial biases, it might be redundant with regards to the already existing continuous number magnitude schema (C0).

Importantly, interindividual variability, as postulated by the Psychological Markedness account and demonstrated at least for MC (Hohol et al., 2020, Supplementary Materials) and PJ (Cipora, van Dijck, et al., 2019), is accommodated in our model through the parameters optimization being run at the individual level. In other words, in contrast to the majority of existing theoretical accounts, our model treats the SNARC and SNARC-like effects as indominant phenomena (Rouder & Haaf, 2018), which they appear to be. Although usually not strived for when searching for general tendencies in cognitive processing, large between-subject variability in optimized parameters reflects empirical observations realistically. Additionally, we may try to account for intraindividual variability in the future (see Study 6) to account for the possibility that a distribution of plausible SNARC effects exists within each participant, rather than a stable effect. For this, we may either supplement the model with longitudinal data from sessions with the same experimental manipulations, in order to capture unsystematic variability, or from sessions that systematically differ in context manipulation. The latter option would account for the claim that psychological markedness of stimulus features systematically varies within individuals. For instance, we may add a cognitive context schema to the model that accounts for influences of the current cultural context, or the language used in the experimental setting within bilingual participants.

The computational contextualization model also includes a schema for ordinality coming from WM (C2), which corresponds to the assumptions of the WM account (van Dijck et al., 2014; van Dijck & Fias, 2011). As of now, this schema only functions for

four different sequences to be memorized, each consisting of four letters (i.e., OJRA, RAOJ, LEWU, and ULEW), as used in the three experiments in Study 4. In future work, a broader range of sequences should be implemented. Notably, a strength of this schema is that, in absence of a newly learned sequence, ordinality within the stimulus set itself is used. This is in line with a revision of the WM account, claiming that a spatial template in WM can also be filled in with elements entailing ordinality stored in LTM (Abrahamse et al., 2016). Moreover, this mechanism may reflect cognitive processes taking place when different ranges of numbers are used as stimuli (see Study 5), where the same numbers may be the first or smallest in one set, but the last or largest in a different set. In other words, the WM ordinality schema (C2) is able to account for relative number magnitude, whereas the continuous, categorical, and small number magnitude contexts (C0, C1, and C4) refer to absolute number magnitude.

Another schema included in the current version of our model reflects cognitive spatial biases specific to letters, which are based on alphabetical ordinality (C3). Similar schemata for other kinds of ordinal stimuli may be implemented in future versions (e.g., a new schema with 12 spatial biases for month names). Extending the current alphabetical ordinality schema to other kinds of ordinal stimuli may be hard, though, because the current schema includes exactly 26 spatial biases (i.e., one per letter position in the alphabet). Please note that, for clarity reasons, Figure 3-3 illustrates only the eight spatial bias distributions corresponding to the letters used in Study 4 and implemented in our current Python code (i.e., A, E, J, L, O, R, U, and W). Notably, the generalization of this schema to other kinds of ordinal stimuli may not even be necessary from a model complexity point of view: Parameters to be optimized or to base data simulation on do not increase with each stimulus-specific schema that is added, because the model deactivates the schemata automatically if not applicable (e.g., C3 in stimuli other than letters; C0, C1, and C4 in stimuli other than numbers).

The additional left-sided bias for small numbers (C4) is related to the idea of a logarithmic MNL (Dehaene, 2001; Izard & Dehaene, 2008; Nuerk, Moeller, et al., 2015), reflected in a higher resolution of the mental representation for small compared to large numbers within the MNL. It may be worthwhile to reconsider this schema so that it can account for different kinds of ordinal stimuli, reflecting left-sided biases for early ordinal positions in any set, instead of left-sided biases specific to small number magnitudes.

Finally, our model also resembles the multiple-coding framework (Schroeder et al., 2017c), in terms of cognitive context schemata functioning similarly to different codes. However, while the multiple-coding framework stays vague on the exact contribution of each postulated code to the SNARC and SNARC-like effects in different tasks, our model is able to estimate the extent to which each schema is activated in different tasks. At present, our model assumes additive contributions of schemata to behavioral patterns. This choice was made to maintain conceptual and computational simplicity in its initial implementation, but we may consider implementing interactions among schemata in the future.

To conclude, the computational contextualization model represents a promising and integrative approach for explaining a variety of experimental findings related to the SNARC and SNARC-like effects. The model combines aspects from multiple existing theoretical accounts, and it builds upon the idea that the SNARC and SNARC-like effects result from the simultaneous activation of different cognitive context schemata. With its straightforward and flexible architecture, the model aims at identifying mechanisms that underlie the cognitive processing of numbers and other ordinal stimuli, and to simulate realistic behavioral patterns. The development of the computational contextualization model is ongoing.

3.4. Limitations and future research directions

Although all six studies presented in this dissertation provide important new insights on SNA, they also reveal some limitations and raise new questions to be addressed in future research. To begin with, the approach to examine the presence of reliable SNARC effects using bootstrap confidence intervals applied in Studies 1, 3, and 6 could be run for Studies 2, 4, and 5, as well. We could then compare proportions of samples showing reliable regular or reversed SNARC effects across different experimental setups. Such a broader overview may stimulate new ideas for explaining why a greater number of participants may demonstrate a reliable reversed SNARC effect in CJ (Study 1) and MC (Hohol et al., 2020, Supplementary Material) than in PJ (Cipora, van Dijck, et al., 2019).

Moreover, the “Ironman paradigm” introduced in Study 6 can shed light on the intraindividual variability of many other cognitive phenomena beyond the SNARC effect. At the same time, reapplying it to the SNARC effect could also be worthwhile for investigating its temporal stability in tasks other than PJ. For instance, fluctuations may be more systematic in MC than in PJ, because the magnitude processing is

necessary in MC but not in PJ, which might make its mental representation more stable.

Next, several methodological improvements could be considered for Study 4. This was the only study involving non-numerical stimuli, namely letters, but it included only A, E, J, L, O, R, U, and W. Letter A appeared to have a very robust association with the left side, independent of WM salience, reflecting a spatial mapping for its ordinal position in the alphabet (i.e., stored in LTM) rather than for its ordinal position in the newly learned sets OJRA or RAOJ (i.e., stored in WM). It remains unclear whether such a robust LTM-based association, not overridden by a WM-based association, is unique to the first letter of the alphabet (i.e., A) or whether it also extends to the second and third letters (i.e., B and C), though potentially weaker. Similarly, letter Z, and potentially even Y and X, may be strongly associated with the right side. Our Tübingen lab is therefore planning to re-run Study 4 with letter sets including the first and last few letters in the alphabet. Implementing a larger number of different stimulus sets would allow testing for such stimulus-specific associations while also increasing generalizability. This could even be done in a within-subjects design by dividing the CVC task into more blocks and having participants memorize a new sequence before each block, as in previous studies (Guida et al., 2015; van Dijck & Fias, 2011). Note that in our current study with month names, we could not find specifically robust spatial LTM-based associations for the first and last month in the year (Albu, Scheiderer, et al., 2025).

Many more interesting aspects could be investigated by re-analyzing the data we collected, which is publicly available including documentation (see Table 1-1). For example, potential congruency sequence effects could be explored with the CJ and PJ data gathered in Studies 1, 2, 3, 5, and 6. While beyond the scope of the current dissertation, it would be interesting to find out whether the SNARC effect appears different when considering post-SNARC-congruent trials (i.e., trials following trials in which a small/large number is responded to with the left/right, respectively) compared to post-SNARC-incongruent trials (i.e., trials following trials in which a small/large number is responded to with the right/left, respectively). A large body of research on cognitive control suggests that the size of congruency effects is typically smaller following incongruent than congruent trials (e.g., Egner, 2007). If we conceptualize the SNARC effect as a compatibility effect, we would predict it to be stronger when considering only post-SNARC-congruent as opposed to post-SNARC-incongruent

trials. Note, however, that some participants show a reversed overall SNARC effect, which must be taken into account when defining congruent and incongruent trials. Another example of an interesting trial-by-trial phenomenon worth exploring in our data, and also related to cognitive control, is post-error slowing (and also the related phenomenon of pre-error speeding; e.g., Pfister & Foerster, 2022). On the one hand, the SNARC effect might be stronger in trials following those in which an error was made, because such trials are typically responded to more slowly, and the SNARC effect is generally more pronounced in slower responses. On the other hand, enhanced cognitive control in these trials might suppress the activation of task-irrelevant associations, such as the mental representation of number magnitude. Importantly, anyone can investigate the SNARC effect in relation to classic response-time patterns, (e.g., congruency sequence effects, post-error slowing), because we have made the data from all our studies publicly available.

Finally, the newly introduced computational contextualization model seems to be a useful approach to determining cognitive mechanisms that underlie the SNARC and SNARC-like effects. As outlined in Section 3.3.3., several additional steps are planned to further enhance the model.

4. Conclusion

This dissertation has addressed several controversies in numerical cognition by investigating the spatial mental representations of number magnitude and ordinality across six studies. Insights were drawn from different tasks, cultural contexts, stimuli, and time points, primarily through large online experiments. The findings indicate that the processing of magnitude and ordinality, and their spatial mappings, are highly automatic, as shown across a variety of tasks. At the same time, the shape of the SNARC effect depends on task instructions: associations between small and large numbers and left or right side are sometimes continuous, reflecting the underlying structure of number magnitude and ordinality, but at other times categorical. Spatial mappings also appear more robust for small single-digit numbers than for larger ones. At the group level, the strength of the SNARC effect seems to align with language script direction and other cultural directional preferences, although no fully reversed SNARC effect was observed, even in a sample using a right-to-left language script. Beyond numbers, spatial mental representations extend to other ordinal stimuli such as letters, where those of canonical ordinality stored in LTM can coexist with those of arbitrary, newly memorized ordinality stored in WM. Mental representations appear to adapt to the currently relevant stimulus set, but only partially. Crucially, inter- and intraindividual variability of the SNARC effect was high, suggesting that the effect is indominant and temporally unstable. To conclude, this dissertation may be described as a “Bohemian SNARCsody”: Although spatial mental representations of number magnitude and ordinality appear highly automatic and robust at the group level, their precise shape and strength may vary across tasks, cultures, stimuli, and time points. Together with first insights from the computational contextualization model that we are currently developing to account for many of these influences, these findings offer important insights into basic numerical cognition.

References

Abrahamse, E. L., & van Dijck, J.-P. (2023). Ranking-space: Magnitude makes sense through spatially scaffolded ranking. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1224254

Abrahamse, E. L., van Dijck, J.-P., & Fias, W. (2016). How does working memory enable number-induced spatial biases? *Frontiers in Psychology*, *7*, 977. https://doi.org/10.3389/fpsyg.2016.00977

Albu, E., Scheiderer, J., Roth, L., Cipora, K., & Nuerk, H.-C. (2025). *Investigating long- and short-term associations between months and space*. Open Science Framework [Preregistration]. https://doi.org/10.17605/OSF.IO/EBM6F

Albu, E., Slawinski, P., Roth, L., Cipora, K., & Nuerk, H.-C. (2025). *How does a direct task influence the coexistence of short- and long-term associations between letters and space?* Open Science Framework [Preregistration]. https://doi.org/10.17605/OSF.IO/B2Z9M

Anthropic. (2025). *Claude AI* (Version 4 Sonnet) [Large language model]. https://claude.ai

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

Bächtold, D., Baumüller, M., & Brugger, P. (1998). Stimulus-response compatibility in representational space. *Neuropsychologia*, *36*(8), 731–735. https://doi.org/10.1016/S0028-3932(98)00002-5

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In *Handbook of social cognition: Basic processes; Applications* (2nd ed., pp. 1–40). Lawrence Erlbaum Associates, Inc.

Beran, M. J., French, K., Smith, T. R., & Parrish, A. E. (2019). Limited evidence of number–space mapping in rhesus monkeys (Macaca mulatta) and capuchin monkeys (Sapajus apella). *Journal of Comparative Psychology*, *133*(3), 281–293.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, *55*(1), 803–832. https://doi.org/10.1146/annurev.psych.55.090902.141601

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Brysbaert, M. (1995). Arabic number reading: On the nature of the numerical scale and the origin of phonological recoding. *Journal of Experimental Psychology: General*, *124*(4), 434–452. https://doi.org/10.1037/0096-3445.124.4.434

Bueti, D., & Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1525), 1831–1840. https://doi.org/10.1098/rstb.2009.0028

Bulut, M., Candemir, A., Şefikoğlu, M., Haugen, B., Çetinkaya, H., Dural, S., & Dural, S. (2025). *Measuring SNARC effect: Different task setups reveal divergent spatial-numerical associations* [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/2wp9q_v1

Bulut, M., Hepdarcan, I., Palaz, E., Çetinkaya, H., & Dural, S. (2023). No SNARC effect among left-to-right readers: Evidence from a Turkish sample. *Advances in Cognitive Psychology*. https://gcris.ieu.edu.tr/handle/20.500.14365/4792

Bulut, M., Roth, L., Bahreini, N., Cipora, K., Reips, U. D., & Nuerk, H.-C. (2024). One direction? Cultural aspects of the mental number line beyond reading direction. *Psychological Research*, *89*(1), 37. https://doi.org/10.1007/s00426-024-02038-4

Butz, M. V., Mittenbühler, M., Schwöbel, S., Achimova, A., Gumbsch, C., Otte, S., & Kiebel, S. (2025). Contextualizing predictive minds. *Neuroscience & Biobehavioral Reviews*, *168*, 105948. https://doi.org/10.1016/j.neubiorev.2024.105948

Caruso, J. C. (2006). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment*. https://econtent.hogrefe.com/doi/10.1027/1015-5759.20.3.166

Casasanto, D., & Pitt, B. (2019). The faulty magnitude detector: Why SNARC-like tasks cannot support a generalized magnitude system. *Cognitive Science*, *43*(10), 12794. https://doi.org/10.1111/cogs.12794

Cipora, K. (2014). *Between-task consistency of the SNARC effect* [Poster presentation]. XXXIInd European Workshop on Cognitive Neuropsychology, Bressanone, Italy. https://osf.io/nx6vq/

Cipora, K., & Göbel, S. M. (2013). *Number–Space associations: Just how reliable the SNARC effect is* [Poster presentation]. XXXIst European Workshop at Cognitive Neuropsychology, Bressanone, Italy. https://osf.io/nx6vq/

Cipora, K., Haman, M., Domahs, F., & Nuerk, H.-C. (2020). Editorial: On the development of space-number relations: Linguistic and cognitive determinants, influences, and associations. *Frontiers in Psychology*, *11:182*. https://doi.org/10.3389/fpsyg.2020.00182

Cipora, K., He, Y., & Nuerk, H.-C. (2020). The spatial–numerical association of response codes effect and math skills: Why related? *Annals of the New York Academy of Sciences*, *1477*(1), 5–19. https://doi.org/10.1111/nyas.14355

Cipora, K., Hohol, M., Grabner, R. H., Meier, M. A., Roth, L., Sroka, M., & Obidziński, M. (2024). *Professional mathematicians – automaticity*. Open Science Framework [Preregistration]. https://doi.org/10.17605/OSF.IO/WFYQD

Cipora, K., Hohol, M., Nuerk, H.-C., Willmes, K., Brożek, B., Kucharzyk, B., & Nęcka, E. (2016). Professional mathematicians differ from controls in their spatial-numerical associations. *Psychological Research*, *80*(4), 710–726. https://doi.org/10.1007/s00426-015-0677-6

Cipora, K., Patro, K., & Nuerk, H.-C. (2018). Situated influences on Spatial–Numerical Associations. In T. L. Hubbard (Ed.), *Spatial Biases in Perception and Cognition* (pp. 41–59). Cambridge University Press. https://doi.org/10.1017/9781316651247.004

Cipora, K., Roth, L., Rossi, S., Gilmore, C., & Inglis, M. (2023). *Time after time—The Ironman follow-up*. Open Science Framework [Preregistration]. https://doi.org/10.17605/OSF.IO/74MUE

Cipora, K., Soltanlou, M., Reips, U.-D., & Nuerk, H.-C. (2019). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behavior Research Methods*, *51*, 1676–1692. https://doi.org/10.3758/s13428-019-01213-5

Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Goebel, S. M., Willmes, K., Pesenti, M., Schiltz, C., & Nuerk, H.-C. (2019). *A minority pulls the sample mean: On the individual prevalence of robust group-level cognitive phenomena – the instance of the SNARC effect* [Preprint]. Open Science Framework. https://doi.org/10.31234/osf.io/bwyr3

Cipora, K., & Wood, G. (2017). Finding the SNARC instead of hunting it: A 20∗20 Monte Carlo investigation. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01194

Claeys, K. G., Dupont, P., Cornette, L., Sunaert, S., van Hecke, P., Schutter, E., & Orban, G. A. (2004). Color discrimination involves ventral and dorsal stream visual areas. *Cerebral Cortex*, *14*(7), 803–822. https://doi.org/10.1093/cercor/bhh040

Cleland, A. A., & Bull, R. (2019). Automaticity of access to numerical magnitude and its spatial associations: The role of task and number representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(2), 333–348. https://doi.org/10.1037/xlm0000590

Cleland, A. A., & Bull, R. (2025). *SNARC effects for color decision to non-symbolic representations of quantity* [Preprint]. Open Science Framework. https://doi.org/10.31234/osf.io/4za97_v2

Cleland, A. A., Corsico, K., White, K., & Bull, R. (2020). Non-symbolic numerosities do not automatically activate spatial-numerical associations: Evidence from the SNARC effect. *Quarterly Journal of Experimental Psychology*, *73*(2), 295–308. https://doi.org/10.1177/1747021819875021

Cohen Kadosh, R., Cohen Kadosh, K., & Henik, A. (2008). When brightness counts: The neuronal correlate of numerical-luminance interference. *Cerebral Cortex*, *18*(2), 337–343. https://doi.org/10.1093/cercor/bhm058

Cohen Kadosh, R., Henik, A., Rubinsten, O., Mohr, H., Dori, H., van de Ven, V., Zorzi, M., Hendler, T., Goebel, R., & Linden, D. E. J. (2005). Are numbers special? The comparison systems of the human brain investigated by fMRI. *Neuropsychologia*, *43*(9), 1238–1248. https://doi.org/10.1016/j.neuropsychologia.2004.12.017

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*(1), 68–80. https://doi.org/10.1037/h0029382

Crowder, R. G. (1976). *Principles of Learning and Memory*. Psychology Press. https://doi.org/10.4324/9781315746944

Crüwell, S., Doorn, J. van, Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2019). Seven easy steps to Open Science. *Zeitschrift Für Psychologie*. https://econtent.hogrefe.com/doi/10.1027/2151-2604/a000387

Dalmaso, M., Pileggi, S., & Vicovaro, M. (2023). Face age is mapped into three-dimensional space. *Cognitive Science*, *47*(11), e13374. https://doi.org/10.1111/cogs.13374

Dalmaso, M., & Vicovaro, M. (2019). Evidence of SQUARC and distance effects in a weight comparison task. *Cognitive Processing*, *20*(2), 163–173. https://doi.org/10.1007/s10339-019-00905-2

de Hevia, M. D., Girelli, L., Addabbo, M., & Cassia, V. M. (2014). Human infants' preference for left-to-right oriented increasing numerical sequences. *PLOS ONE*, *9*(5), e96412. https://doi.org/10.1371/journal.pone.0096412

de Hevia, M. D., Veggiotti, L., Streri, A., & Bonn, C. D. (2017). At birth, humans associate "few" with left and "many" with right. *Current Biology*, *27*(24), 3879-3884.e2. https://doi.org/10.1016/j.cub.2017.11.024

De Jong, R., Liang, C.-C., & Lauber, E. (1994). Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus-response correspondence. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(4), 731.

Dehaene, S. (2001). Subtracting pigeons: Logarithmic or linear? *Psychological Science*, *12*(3), 244–246. https://doi.org/10.1111/1467-9280.00343

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, *7*(4), 145–147. https://doi.org/10.1016/S1364-6613(03)00055-X

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371–396. https://doi.org/10.1037/0096-3445.122.3.371

Deng, Z., Chen, Y., Zhang, M., Li, Y., & Zhu, X. (2018). The association of number and space under different tasks: Insight from a process perspective. *Frontiers in Psychology*, *9*, 957. https://doi.org/10.3389/fpsyg.2018.00957

Deng, Z., Chen, Y., Zhu, X., & Li, Y. (2017). The effect of working memory load on the SNARC effect: Maybe tasks have a word to say. *Memory & Cognition*, *45*(3), 428–441. https://doi.org/10.3758/s13421-016-0676-x

di Giorgio, E., Lunghi, M., Rugani, R., Regolin, L., Dalla Barba, B., Vallortigara, G., & Simion, F. (2019). A mental number line in human newborns. *Developmental Science*, *22*(6), e12801. https://doi.org/10.1111/desc.12801

Didino, D., Brandtner, M., Glaser, M., & Knops, A. (2023). Probing the dual-route model of the SNARC effect by orthogonalizing processing speed and depth. *Experimental Psychology*. https://econtent.hogrefe.com/doi/10.1027/1618-3169/a000577

Didino, D., Breil, C., & Knops, A. (2019). The influence of semantic processing and response latency on the SNARC effect. *Acta Psychologica*, *196*, 75–86. https://doi.org/10.1016/j.actpsy.2019.04.008

Drucker, C. B., & Brannon, E. M. (2014). Rhesus monkeys (Macaca mulatta) map number onto space. *Cognition*, *132*(1), 57–67. https://doi.org/10.1016/j.cognition.2014.03.011

Dural, S., Çetinkaya, H., Hepdarcan, I., Gür, E., & Korkut, İ. (2025). Revisiting the SNARC effect: Testing magnitude classification in a Turkish sample typically lacking the SNARC effect. *Journal of Cognitive Psychology*, 1–10. https://doi.org/10.1080/20445911.2025.2482600

Egner, T. (2007). Congruency sequence effects and cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 380–390. https://doi.org/10.3758/CABN.7.4.380

Endler, N. S., & Kocovski, N. L. (2001). State and trait anxiety revisited. *Journal of Anxiety Disorders*, *15*(3), 231–245. https://doi.org/10.1016/S0887-6185(01)00060-3

Erfanikia, M. (2024). *Is face age mapped horizontally into space in a culture with mixed reading habits? A cross-cultural study* [Master's thesis]. University of Padova. https://thesis.unipd.it/handle/20.500.12608/69961

Faghihi, N., & Vaid, J. (2023). Reading/writing direction as a source of directional bias in spatial cognition: Possible mechanisms and scope. *Psychonomic Bulletin & Review*, *30*(3), 843–862. https://doi.org/10.3758/s13423-022-02239-1

Farrell, S., & Lewandowsky, S. (n.d.). *Computational Modeling of Cognition and Behavior*.

Feldman, A., Oscar-Strom, Y., Tzelgov, J., & Berger, A. (2019). Spatial–numerical association of response code effect as a window to mental representation of magnitude in long-term memory among Hebrew-speaking children. *Journal of Experimental Child Psychology*, *181*, 102–109. https://doi.org/10.1016/j.jecp.2019.01.001

Felisatti, A., Aagten-Murphy, D., Laubrock, J., Shaki, S., & Fischer, M. H. (2020). The brain's asymmetric frequency tuning: Asymmetric behavior originates from asymmetric perception. *Symmetry*, *12*, Article 12. https://doi.org/10.3390/sym12122083

Felisatti, A., Laubrock, J., Shaki, S., & Fischer, M. H. (2020a). A biological foundation for spatial–numerical associations: The brain's asymmetric frequency tuning. *Annals of the New York Academy of Sciences*, *1477*, 44–53. https://doi.org/10.1111/nyas.14418

Felisatti, A., Laubrock, J., Shaki, S., & Fischer, M. H. (2020b). Commentary: A mental number line in human newborns. *Frontiers in Human Neuroscience*, *14*, Article 99. https://doi.org/10.3389/fnhum.2020.00099

Fias, W., Brysbaert, M., Geypens, F., & D'Ydewalle, G. (1996). The importance of magnitude information in numerical processing: Evidence from the SNARC effect. *Mathematical Cognition*, *2*(1), 95–110. https://doi.org/10.1080/135467996387552

Fias, W., Lauwereyns, J., & Lammertyn, J. (2001). Irrelevant digits affect feature-based attention depending on the overlap of neural circuits. *Cognitive Brain Research*, *12*(3), 415–423. https://doi.org/10.1016/S0926-6410(01)00078-7

Fias, W., & van Dijck, J.-P. (2016). The temporary nature of number-space interactions. *Canadian Journal of Experimental Psychology*, *70*(1), 33–40. https://doi.org/10.1037/cep0000071

Fischer, M. H., Mills, R. A., & Shaki, S. (2010). How to cook a SNARC: number placement in text rapidly changes spatial-numerical associations. *Brain and Cognition*, *72*(3), 333–336. https://doi.org/10.1016/j.bandc.2009.10.010

Fischer, M. H., Shaki, S., & Cruise, A. (2009). It takes just one word to quash a SNARC. *Experimental Psychology*, *56*(5), 361–366. https://doi.org/10.1027/1618-3169.56.5.361

Foucault, M. (1928). Les inhibitions internes de fixation. *L'Année psychologique*, *29*(1), 92–112. https://doi.org/10.3406/psy.1928.4805

Ftaïta, M., Guida, A., Fartoukh, M., & Mathy, F. (2024). Spatial–positional associations in short-term memory can vanish in long-term memory. *Memory & Cognition*. https://doi.org/10.3758/s13421-024-01577-w

Fum, D., Missier, F. D., & Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research*, *8*(3), 135–142. https://doi.org/10.1016/j.cogsys.2007.07.001

Galton, F. (1880). Visualised Numerals. *Nature*, *21*(533), 252–256. https://doi.org/10.1038/021252a0

Garaizar, P., & Reips, U.-D. (2019). Best practices: Two Web-browser-based methods for stimulus presentation in behavioral experiments with high-resolution timing

requirements. *Behavior Research Methods*, *51*(3), 1441–1453. https://doi.org/10.3758/s13428-018-1126-4

Gazes, R. P., Diamond, R. F. L., Hope, J. M., Caillaud, D., Stoinski, T. S., & Hampton, R. R. (2017). Spatial representation of magnitude in gorillas and orangutans. *Cognition*, *168*, 312–319. https://doi.org/10.1016/j.cognition.2017.07.010

Georges, C., Hoffmann, D., & Schiltz, C. (2013). *The SNARC effect and its relationship to spatial abilities in women* [Poster presentation]. 18th conference of the European Society for Cognitive Psychology, Budapest, Hungary. https://orbilu.uni.lu/handle/10993/13012

Gevers, W., Ratinckx, E., Baene, W., & Fias, W. (2006). Further evidence that the SNARC effect is processed along a dual-route architecture: Evidence from the lateralized readiness potential. *Experimental Psychology*, *53*(1), 58–68. https://doi.org/10.1027/1618-3169.53.1.58

Gevers, W., Reynvoet, B., & Fias, W. (2003). The mental representation of ordinal sequences is spatially organized. *Cognition*, *87*(3), B87–B95. https://doi.org/10.1016/S0010-0277(02)00234-2

Gevers, W., Reynvoet, B., & Fias, W. (2004). The mental representation of ordinal sequences is spatially organised: Evidence from days of the week. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*. https://psycnet.apa.org/record/2004-12531-025

Gevers, W., Santens, S., Dhooge, E., Chen, Q., van den Bossche, L., Fias, W., & Verguts, T. (2010). Verbal-spatial and visuospatial coding of number-space interactions. *Journal of Experimental Psychology: General*, *139*(1), 180–190. https://doi.org/10.1037/a0017688

Gevers, W., Verguts, T., Reynvoet, B., Caessens, B., & Fias, W. (2006). Numbers and space: A computational model of the SNARC effect. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 32–44. https://doi.org/10.1037/0096-1523.32.1.32

Ginsburg, V., & Gevers, W. (2015). Spatial coding of ordinal information in short- and long-term memory. *Frontiers in Human Neuroscience*, *9*(8). https://doi.org/10.3389/fnhum.2015.00008

Ginsburg, V., van Dijck, J.-P., Previtali, P., Fias, W., & Gevers, W. (2014). The impact of verbal working memory on number–space associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 976–986. https://doi.org/10.1037/a0036378

Giuliani, F., Brunello, L., Dalmaso, M., D'Anselmo, A., Tommasi, L., & Vicovaro, M. (2024). The right side of price: Evidence of a SNARC-like effect for economic value. *Current Psychology*, *43*(20), 18330–18343. https://doi.org/10.1007/s12144-024-05612-6

Giurfa, M., Marcout, C., Hilpert, P., Thevenot, C., & Rugani, R. (2022). An insect brain organizes numbers on a left-to-right mental number line. *Proceedings of the National Academy of Sciences*, *119*(44), Article e2203584119. https://doi.org/10.1073/pnas.2203584119

Gökaydin, D., Brugger, P., & Loetscher, T. (2018). Sequential Effects in SNARC. *Scientific Reports*, *8*(1), 10996. https://doi.org/10.1038/s41598-018-29337-2

Guida, A., & Campitelli, G. (2019). Explaining the SPoARC and SNARC effects with knowledge structures: An expertise account. *Psychonomic Bulletin & Review*, *26*(2), 434–451. https://doi.org/10.3758/s13423-019-01582-0

Guida, A., & Lavielle-Guida, M. (2014). 2011 space odyssey: Spatialization as a mechanism to code order allows a close encounter between memory expertise and classic immediate memory studies. *Frontiers in Psychology*, *5*, Article 573. https://doi.org/10.3389/fpsyg.2014.00573

Guida, A., Leroux, A., Lavielle-Guida, M., & Noël, Y. (2015). A SPoARC in the dark: Spatialization in verbal immediate memory. *Cognitive Science*, *40*(8), 2018–2121. https://doi.org/10.1111/cogs.12316

Guida, A., Megreya, A. M., Lavielle-Guida, M., Noël, Y., Mathy, F., van Dijck, J.-P., & Abrahamse, E. L. (2018). Spatialization in working memory is related to literacy and

reading direction: Culture “literarily” directs our thoughts. *Cognition*, *175*, 96–100. https://doi.org/10.1016/j.cognition.2018.02.013

Guida, A., Mosinski, F., Cipora, K., Mathy, F., & Noël, Y. (2020). Spatialization in working memory: Can individuals reverse the cultural direction of their thoughts? *Annals of the New York Academy of Sciences*, *1477*(1), 113–125. https://doi.org/10.1111/nyas.14499

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don’t? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*(3), 772–789. https://doi.org/10.3758/s13423-018-1522-x

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Haspelmath, M. (2006). Against markedness (and what to replace it with). *Journal of Linguistics*, *42*(1), 25–70. https://doi.org/10.1017/S0022226705003683

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Herrera, A., Macizo, P., & Semenza, C. (2008). The role of working memory in the association between number magnitude and space. *Acta Psychologica*, *128*(2), 225–237. https://doi.org/10.1016/j.actpsy.2008.01.002

Hesse, P. N., & Bremmer, F. (2017). The SNARC effect in two dimensions: Evidence for a frontoparallel mental number plane. *Vision Research*, *130*, 85–96. https://doi.org/10.1016/j.visres.2016.10.007

Heubner, L., Cipora, K., Soltanlou, M., Schlenker, M.-L., Lipowska, K., Göbel, S. M., Domahs, F., Haman, M., & Nuerk, H.-C. (2018). A mental odd-even continuum account: Some

numbers may be “more odd” than others and some numbers may be “more even” than others. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01081

Hines, T. M. (1990). An odd effect: Lengthened reaction times for judgments about odd digits. *Memory & Cognition*, *18*(1), 40–46. https://doi.org/10.3758/BF03202644

Hines, T. M. (2010). The Odd Effect and Multiple Meanings in English of the Words “Odd” and “Even.” *Psychological Reports*, *107*(2), 443–446. https://doi.org/10.2466/04.10.11.22.PR0.107.5.443-446

Hochman, S., Havedanloo, R., Heysieattalab, S., & Soltanlou, M. (2024). How does language modulate the association between number and space? A registered report of a cross-cultural study of the spatial–numerical association of response codes effect. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001653

Hoffmann, D., Hornung, C., Martin, R., & Schiltz, C. (2013). Developing number-space associations: SNARC effects using a color discrimination task in 5-year-olds. *Journal of Experimental Child Psychology*, *116*(4), 775–791. https://doi.org/10.1016/j.jecp.2013.07.013

Hohol, M., Willmes, K., Nęcka, E., Brożek, B., Nuerk, H.-C., & Cipora, K. (2020). Professional mathematicians do not differ from others in the symbolic numerical distance and size effects. *Scientific Reports*, *10*(1), 11531. https://doi.org/10.1038/s41598-020-68202-z

Huang, K. (2024). Eye-tracking evidence for the SNARC effect: Unveiling attentional stages. *Interdisciplinary Humanities and Communication Studies*, *1*(9), Article 9. https://doi.org/10.61173/fd6be159

Huber, S., Klein, E., Graf, M., Nuerk, H.-C., Moeller, K., & Willmes, K. (2015). Embodied markedness of parity? Examining handedness effects on parity judgments. *Psychological Research*, *79*(6), 963–977. https://doi.org/10.1007/s00426-014-0626-9

Huber, S., Klein, E., Moeller, K., & Willmes, K. (2016). Spatial-numerical and ordinal positional associations coexist in parallel. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00438

Hume, E. (2004). Deconstructing Markedness: A predictability-based approach. *Annual Meeting of the Berkeley Linguistics Society*, 182–198. http://journals.linguisticsociety.org/proceedings/index.php/BLS/article/download/948/729

Hung, Y., Hung, D. L., Tzeng, O. J.-L., & Wu, D. H. (2008). Flexible spatial mapping of different notations of numbers in Chinese readers. *Cognition*, *106*(3), 1441–1450. https://doi.org/10.1016/j.cognition.2007.04.017

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221–1247. https://doi.org/10.1016/j.cognition.2007.06.004

Johnson-Ulrich, Z., & Vonk, J. (2018). Spatial representation of magnitude in humans (Homo sapiens), Western lowland gorillas (Gorilla gorilla gorilla), and American black bears (Ursus americanus). *Animal Cognition*, *21*(4), 531–550. https://doi.org/10.1007/s10071-018-1186-y

Judd, N., Aristodemou, M., & Kievit, R. (2023). *Cognitive variability is ubiquitous and distinct from mean performance across eleven tasks with over 7 million trials* [Preprint]. Open Science Framework. https://doi.org/10.31234/osf.io/b29rn

Kaya, C., Candemır, A., Kaya, D., Çetınkaya, H., & Dural, S. (2025). Investigating no SNARC: Do reading habits provide insight into the SNARC patterns of Turkish sample? *Journal of Cognition and Culture*, *25*(3–4), 421–437. https://doi.org/10.1163/15685373-12340218

Keus, I. M. (2006). *Shedding light on the SNARC: Illuminating the association between numerical and spatial information* [Radboud Universiteit Nijmegen, the Netherlands]. https://repository.ubn.ru.nl/handle/2066/56964

Keus, I. M., Jenks, K. M., & Schwarz, W. (2005). Psychophysiological evidence that the SNARC effect has its functional locus in a response selection stage. *Cognitive Brain Research*, *24*(1), 48–56. https://doi.org/10.1016/j.cogbrainres.2004.12.005

Keus, I. M., & Schwarz, W. (2005). Searching for the functional locus of the SNARC effect: Evidence for a response-related origin. *Memory & Cognition*, *33*(4), 681–695. https://doi.org/10.3758/BF03195335

Koch, N. N., Huber, J. F., Lohmann, J., Cipora, K., Butz, M. V., & Nuerk, H.-C. (2023). Mental number representations are spatially mapped both by their magnitudes and ordinal positions. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.67908

Krajcsi, A., Lengyel, G., & Laczkó, Á. (2018). Interference between number magnitude and parity. *Experimental Psychology*, *65*(2), 71–83. https://doi.org/10.1027/1618-3169/a000394

Krantz, J. H., & Reips, U.-D. (2017). The state of web-based research: A survey and call for inclusion in curricula. *Behavior Research Methods*, *49*(5), 1621–1629. https://doi.org/10.3758/s13428-017-0882-x

Lammertyn, J., Fias, W., & Lauwereyns, J. (2002). Semantic influences on feature-based attention due to overlap of neural circuits. *Cortex*, *38*(5), 878–882. https://doi.org/10.1016/S0010-9452(08)70061-3

Lengyel, G., & Krajcsi, A. (2024). *The SNARC effect emerges only with symbolic numbers* [Preprint]. Open Science Framework. https://doi.org/10.31234/osf.io/2c9kr

Lenoir, J., & Badets, A. (2025). Effect of spatial training on space-number mapping: A situated cognition account. *Psychological Research*, *89*(1), 49. https://doi.org/10.1007/s00426-025-02078-4

Löffler, C. S., Gerten, J., Mamporia, M., Müller, J., Neu, T., Rumpf, J., Schiller, M., Schneider, Y., Wozniak, M., & Topolinski, S. (2022). Bright on the right feels right: SQUARC compatibility is hedonically marked. *Cognition & Emotion*, *36*(4), 767–772. https://doi.org/10.1080/02699931.2022.2053660

Lopiccolo, D., & Chang, C. B. (2021). Cultural factors weaken but do not reverse left-to-right spatial biases in numerosity processing: Data from Arabic and English monoliterates and Arabic-English biliterates. *PLOS ONE*, *16*(12), Article e0261146. https://doi.org/10.1371/journal.pone.0261146

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 149–157. https://doi.org/10.1037/0278-7393.16.1.149

Lozin, M., & Pinhas, M. (2024). Perceiving the “smallest” or “largest” multidigit number: A novel numeric-scale end effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *51*(3), 371–390. https://doi.org/10.1037/xlm0001386

Luck, S. (2019, February 20). Why experimentalists should ignore reliability and focus on precision. *Luck Lab*. https://lucklab.ucdavis.edu/blog/2019/2/19/reliability-and-precision

Lyons, I. M., Vogel, S. E., & Ansari, D. (2016). Chapter 8 - On the ordinality of numbers: A review of neural and behavioral studies. In M. Cappelletti & W. Fias (Eds.), *Progress in Brain Research* (Vol. 227, pp. 187–221). Elsevier. https://doi.org/10.1016/bs.pbr.2016.04.010

Macnamara, A., Keage, H. A. D., & Loetscher, T. (2018). Mapping of non-numerical domains on space: A systematic review and meta-analysis. *Experimental Brain Research*, *236*(2), 335–346. https://doi.org/10.1007/s00221-017-5154-6

Manavusak, C. (2025). *Investigating the SNARC effect: Implicit applications and test-retest reliability*. https://thesis.unipd.it/handle/20.500.12608/82427

Masson, N., Andres, M., Alsamour, M., Bollen, Z., & Pesenti, M. (2020). Spatial biases in mental arithmetic are independent of reading/writing habits: Evidence from French and Arabic speakers. *Cognition*, *200*, 104262. https://doi.org/10.1016/j.cognition.2020.104262

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McCrink, K., & de Hevia, M. D. (2018). From innate spatial biases to enculturated spatial cognition: The case of spatial associations in number and other sequences. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00415

McCrink, K., & Opfer, J. E. (2014). Development of Spatial-Numerical Associations. *Current Directions in Psychological Science*, *23*(6), 439–445. https://doi.org/10.1177/0963721414549751

McManus, R. M., Young, L., & Sweetman, J. (2023). Psychology is a property of persons, not averages or distributions: Confronting the group-to-person generalizability problem in experimental psychology. *Advances in Methods and Practices in Psychological Science*, *6*(3). https://doi.org/10.1177/25152459231186615

Mercury, F. (1975). *Bohemian Rhapsody* [Song on the album “A Night at the Opera” recorded by Queen]. EMI Records (UK), Elektra Records (USA), Queen Music Ltd.

Miccoli, M. R., & Reips, U.-D. (2025). Exact versus conceptual replication: Internet-based research investigating the replicability of cognitive effects. *Journal of Perceptual Imaging*, *7*, 1–15. https://doi.org/10.2352/J.Percept.Imaging.2025.7.000408

Microsoft Corporation. (2024). *PowerPoint for Microsoft 365* (Version 2408) [Computer software]. Microsoft. https://www.microsoft.com

Mingolo, S., Prpic, V., Mariconda, A., Brugger, P., Drack, T., Bilotta, E., Agostini, T., & Murgia, M. (2023). It’s SNARC o’ clock: Manipulating the salience of the context in a conceptual replication of Bächtold et al.’s (1998) clockface study. *Psychological Research*, 1–15. https://doi.org/10.1007/s00426-023-01893-x

Mitchell, T., Bull, R., & Cleland, A. A. (2012). Implicit response-irrelevant number information triggers the SNARC effect: Evidence using a neural overlap paradigm. *Quarterly Journal of Experimental Psychology*, *65*(10), 1945–1961. https://doi.org/10.1080/17470218.2012.673631

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology*, *67*, 263–287. https://doi.org/10.1146/annurev-psych-122414-033550

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–1520. https://doi.org/10.1038/2151519a0

Nieder, A. (2016). Representing something out of nothing: The dawning of zero. *Trends in Cognitive Sciences*, *20*(11), 830–842. https://doi.org/10.1016/j.tics.2016.08.008

Nuerk, H.-C., Iversen, W., & Willmes, K. (2004). Notational modulation of the SNARC and the MARC (linguistic markedness of response codes) effect. *Quarterly Journal of Experimental Psychology*, *57*(5), 835–863. https://doi.org/10.1080/02724980343000512

Nuerk, H.-C., Moeller, K., Klein, E., Willmes, K., & Fischer, M. H. (2015). Extending the Mental Number Line. *Zeitschrift Für Psychologie*, *219*(1), 3–22. https://doi.org/10.1027/2151-2604/a000041

Nuerk, H.-C., Patro, K., Cress, U., Schild, U., Friedrich, C. K., & Göbel, S. M. (2015). How space-number associations may be created in preliterate children: Six distinct mechanisms. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00215

Nuerk, H.-C., & Schroeder, P. A. (2024). *A psychological markedness account in numerical cognition* [Unpublished manuscript].

Núñez, R. E. (2011). No innate number line in the human brain. *Journal of Cross-Cultural Psychology*, *42*(4), 651–668. https://doi.org/10.1177/0022022111406097

Opfer, J. E., & Furlong, E. E. (2011). How numbers bias preschoolers' spatial search. *Journal of Cross-Cultural Psychology*, *42*(4), 682–695. https://doi.org/10.1177/0022022111406098

Parrott, J. (2013, April 24). *Hadith on manners: Eating and drinking with right hand*. Daily Hadith Online - The Sunnah of Prophet Muhammad. https://www.abuaminaelias.com/dailyhadithonline/2013/04/24/eat-drink-right-hand/

Parrott, J. (2017, December 1). *Hadith on mindfulness: Beginning actions with the right side*. Daily Hadith Online - The Sunnah of Prophet Muhammad. https://www.abuaminaelias.com/dailyhadithonline/2017/12/01/beginning-deeds-yamin-yameen/

Patro, K., Fischer, U., Nuerk, H.-C., & Cress, U. (2016). How to rapidly construct a spatial–numerical representation in preliterate children (at least temporarily). *Developmental Science*, *19*(1), 126–144. https://doi.org/10.1111/desc.12296

Patro, K., Nuerk, H.-C., & Cress, U. (2016). Mental number line in the preliterate brain: The role of early directional experiences. *Child Development Perspectives*, *10*(3), 172–177. https://doi.org/10.1111/cdep.12179

Patro, K., Nuerk, H.-C., Cress, U., & Haman, M. (2014). How number-space relationships are assessed before formal schooling: A taxonomy proposal. *Frontiers in Psychology*, *5*, 419. https://doi.org/10.3389/fpsyg.2014.00419

Pfister, R., & Foerster, A. (2022). How to measure post-error slowing: The case of pre-error speeding. *Behavior Research Methods*, *54*(1), 435–443. https://doi.org/10.3758/s13428-021-01631-4

Pfister, R., Schroeder, P. A., & Kunde, W. (2013). SNARC struggles: Instant control over spatial-numerical associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1953–1958. https://doi.org/10.1037/a0032991

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron*, *41*(6), 983–993. https://doi.org/10.1016/S0896-6273(04)00107-2

Pinhas, M., Pothos, E. M., & Tzelgov, J. (2013). Zooming in and out from the mental number line: Evidence for a number range effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 972–976. https://doi.org/10.1037/a0029527

Pinto, M., Pellegrino, M., Marson, F., Lasaponara, S., Cestari, V., D'Onofrio, M., & Doricchi, F. (2021). How to trigger and keep stable directional Space-Number Associations (SNAs). *Cortex*, *134*, 253–264. https://doi.org/10.1016/j.cortex.2020.10.020

Pinto, M., Pellegrino, M., Marson, F., Lasaponara, S., & Doricchi, F. (2019). Reconstructing the origins of the space-number association: Spatial and number-magnitude codes must be used jointly to elicit spatially organised mental number lines. *Cognition*, *190*, 143–156. https://doi.org/10.1016/j.cognition.2019.04.032

Pitt, B., & Casasanto, D. (2022). The order of magnitude: Why SNARC-like tasks (still) cannot support a generalized magnitude system. *Cognitive Science*, *46*(2), Article 13108. https://doi.org/10.1111/cogs.13108

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology*, *17*(2), 132–138. https://doi.org/10.1080/17470216508416422

Previtali, P., de Hevia, M. D., & Girelli, L. (2010). Placing order in space: The SNARC effect in serial learning. *Experimental Brain Research*, *201*(3), 599–605. https://doi.org/10.1007/s00221-009-2063-3

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, *132*(3), 416–442. https://doi.org/10.1037/0033-2909.132.3.416

Prpic, V., Fumarola, A., Tommaso, M., Luccio, R., Murgia, M., & Agostini, T. (2016). Separate mechanisms for magnitude and order processing in the spatial-numerical association of response codes (SNARC) effect: The strange case of musical note values. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1241–1251. https://doi.org/10.1037/xhp0000217

Prpic, V., Mingolo, S., Agostini, T., & Murgia, M. (2021). Magnitude and order are both relevant in SNARC and SNARC-like effects: A commentary on Casasanto and Pitt (2019). *Cognitive Science*, *45*(7), Article 13006. https://doi.org/10.1111/cogs.13006

Python Software Foundation. (2025). *Python Language* (Version 3.12.9) [Computer software]. https://www.python.org/

R Core Team. (2025). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rasoulzadeh, V., van Dijck, J.-P., Khosrowabadi, R., & Fias, W. (2024). Reading direction interacts with spatial processes of temporal order verbal working memory: Evidence from Iranian right-to-left readers. *Language, Cognition and Neuroscience*, *39*(8), 1072–1089. https://doi.org/10.1080/23273798.2024.2370046

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–117). Academic Press. https://doi.org/10.1016/B978-012099980-4/50005-8

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. https://doi.org/10.1026//1618-3169.49.4.243

Reips, U.-D. (2007). The methodology of Internet-based experiments. In A. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford Handbook of Internet Psychology* (pp. 373–390). Oxford University Press. 10.1093/oxfordhb/9780199561803.013.0024

Reips, U.-D. (2009). Internet experiments: Methods, guidelines, metadata. In B. E. Rogowitz & T. N. Pappas (Eds.), *Human Vision and Electronic Imaging XIV* (p. 724008). International Society for Optics and Photonics (SPIE). https://doi.org/10.1117/12.823416

Reips, U.-D. (2021). Web-based research in psychology. *Zeitschrift Für Psychologie*, *229*(4), 198–213. https://doi.org/10.1027/2151-2604/a000475

Reips, U.-D., & Neuhaus, C. (2002). WEXTOR: A Web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, & Computers*, *34*(2), 234–240. https://doi.org/10.3758/bf03195449

Ren, P., Nicholls, M. E. R., Ma, Y., & Chen, L. (2011). Size matters: Non-numerical magnitude affects the spatial coding of response. *PLOS ONE*, *6*(8), Article 23553. https://doi.org/10.1371/journal.pone.0023553

Restle, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology*, *83*(2, Pt.1), 274–278. https://doi.org/10.1037/h0028573

Rice, K. (2007). Markedness in phonology. In P. de Lacy (Ed.), *The Cambridge Handbook of Phonology* (pp. 79–97). Cambridge University Press. https://doi.org/10.1017/CBO9780511486371.005

Richter, M., & Wühr, P. (2022). The nature of associations between physical stimulus size and left-right response codes. *Journal of Cognition*, *5*(1), 15. https://doi.org/10.5334/joc.206

Riello, M., & Rusconi, E. (2011). Unimanual SNARC effect: Hand matters. *Frontiers in Psychology*, *2*, Article 372. https://doi.org/10.3389/fpsyg.2011.00372

Roettger, T. B., & Domahs, F. (2015). Grammatical number elicits SNARC and MARC effects as a function of task demands. *Quarterly Journal of Experimental Psychology*, *68*(6), 1231–1248. https://doi.org/10.1080/17470218.2014.979843

Ross-Hellauer, T. (2022). Open science, done wrong, will compound inequities. *Nature*, *603*(7901), 363–363. https://doi.org/10.1038/d41586-022-00724-0

Roth, L., Findik, Y., Dilek, H., Bulut, M., Cipora, K., & Nuerk, H.-C. (2025). *Influence of cultural factors and context on the SNARC effect in German-Turkish bilinguals*. Open Science Framework [Preregistration]. https://doi.org/10.17605/OSF.IO/VCKMQ

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19–26. https://doi.org/10.1177/2515245917745058

Rouder, J. N., Mehrvarz, M., & Schnuerch, M. (2025). *The role of reliability in experiments* [Preprint]. Open Science Framework. https://doi.org/10.31234/osf.io/xfpmb_v1

Rugani, R., & de Hevia, M.-D. (2017). Number-space associations without language: Evidence from preverbal human infants and non-human animal species. *Psychonomic Bulletin & Review*, *24*(2), 352–369. https://doi.org/10.3758/s13423-016-1126-2

Rugani, R., Kelly, D. M., Szelest, I., Regolin, L., & Vallortigara, G. (2010). Is it only humans that count from left to right? *Biology Letters*, *6*(3), 290–292. https://doi.org/10.1098/rsbl.2009.0960

Rugani, R., Vallortigara, G., Priftis, K., & Regolin, L. (2015). Number-space mapping in the newborn chick resembles humans' mental number line. *Science*, *347*(6221), 534–536.

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017a). Prefrontal neuromodulation reverses spatial associations of non-numerical sequences, but not numbers. *Biological Psychology*, *128*, 39–49. https://doi.org/10.1016/j.biopsycho.2017.07.008

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017b). Space in numerical and ordinal information: A common construct? *Journal of Numerical Cognition*, *3*(2), 164–181. https://doi.org/10.5964/jnc.v3i2.40

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017c). Switching between multiple codes of SNARC-like associations: Two conceptual replication attempts with anodal tDCS in sham-controlled cross-over design. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00654

Schuch, S., Philipp, A. M., Maulitz, L., & Koch, I. (2022). On the reliability of behavioral measures of cognitive control: Retest reliability of task-inhibition effect, task-preparation effect, Stroop-like interference, and conflict adaptation effect. *Psychological Research*, *86*(7), 2158–2184. https://doi.org/10.1007/s00426-021-01627-x

Schwarz, W., & Keus, I. M. (2004). Moving the eyes along the mental number line: Comparing SNARC effects with saccadic and manual responses. *Perception & Psychophysics*, *66*(4), 651–664. https://doi.org/10.3758/BF03194909

Schwarz, W., & Müller, D. (2006). Spatial associations in number-related tasks: A comparison of manual and pedal responses. *Experimental Psychology*, *53*(1), 4–15. https://doi.org/10.1027/1618-3169.53.1.4

Shaki, S., & Fischer, M. H. (2008). Reading space into numbers: A cross-linguistic comparison of the SNARC effect. *Cognition*, *108*(2), 590–599. https://doi.org/10.1016/j.cognition.2008.04.001

Shaki, S., & Fischer, M. H. (2018). Deconstructing spatial-numerical associations. *Cognition*, *175*, 109–113. https://doi.org/10.1016/j.cognition.2018.02.022

Shaki, S., Fischer, M. H., & Göbel, S. M. (2012). Direction counts: A comparative study of spatially directional counting biases in cultures with different reading directions. *Journal of Experimental Child Psychology*, *112*(2), 275–281. https://doi.org/10.1016/j.jecp.2011.12.005

Shaki, S., Fischer, M. H., & Petrusic, W. M. (2009). Reading habits for both words and numbers contribute to the SNARC effect. *Psychonomic Bulletin & Review*, *16*(2), 328–331. https://doi.org/10.3758/PBR.16.2.328

Shen, J., He, H., Wu, B., & Zhou, J. (2023). Influence of different spatial representations on the SNARC effect for letters: Electrophysiological evidence. *Quarterly Journal of Experimental Psychology*, *76*(11), 2613–2628. https://doi.org/10.1177/17470218231167056

Silva, A. A., Pitt, B., & Ferrigno, S. (2025). Rhesus monkeys show no preference for a left-to-right number-space mapping. *Proceedings of the 47th Annual Conference of the Cognitive Science Society*, *47*, 2931–2937. https://escholarship.org/uc/item/3cm2j0gd

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Spaepen, E., Gunderson, E. A., Gibson, D., Goldin-Meadow, S., & Levine, S. C. (2018). Meaning before order: Cardinal principle knowledge predicts improvement in understanding the successor principle and exact ordering. *Cognition*, *180*, 59–81. https://doi.org/10.1016/j.cognition.2018.06.012

Spellman, B., Gilbert, E., & Corker, K. S. (2017). *Open Science: What, why, and how* [Preprint]. Open Science Framework. https://doi.org/10.31234/osf.io/ak6jr

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652–654. https://doi.org/10.1126/science.153.3736.652

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457.

Sternberg, S. (2016). In defence of high-speed memory scanning. *Quarterly Journal of Experimental Psychology*, *69*(10), 2020–2075. https://doi.org/10.1080/17470218.2016.1198820

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181. https://doi.org/10.1037/h0046162

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389–408. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

The pandas development team. (2020). *Pandas 1.0.3* [Computer software]. https://doi.org/10.5281/zenodo.3715232

Toomarian, E. Y., & Hubbard, E. M. (2018). On the genesis of Spatial-Numerical Associations: Evolutionary and cultural factors co-construct the Mental Number Line. *Neuroscience & Biobehavioral Reviews*, *90*, 184–199. https://doi.org/10.1016/j.neubiorev.2018.04.010

Treisman, A., Vieira, A., & Hayes, A. (1992). Automaticity and preattentive processing. *The American Journal of Psychology*, *105*(2), 341–362. https://doi.org/10.2307/1423032

Triki, Z., & Bshary, R. (2018). Cleaner fish Labroides dimidiatus discriminate numbers but fail a mental number line test. *Animal Cognition*, *21*(1), 99–107. https://doi.org/10.1007/s10071-017-1143-1

Tzelgov, J., Ganor-Stern, D., Kallai, A. Y., & Pinhas, M. (2015). Primitives and non-primitives of numerical representations. In R. C. Kadosh & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 45–66). Oxford University Press.

Tzelgov, J., Meyer, J., & Henik, A. (1992). Automatic and intentional processing of numerical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(1), 166–179. https://doi.org/10.1037/0278-7393.18.1.166

Tzelgov, J., Zohar-Shai, B., & Nuerk, H.-C. (2013). On defining, quantifying, and measuring the SNARC effect. *Frontiers in Psychology*, *4*, 302. https://doi.org/10.3389/fpsyg.2013.00302

Ulrich, R. (2009). Uncovering unobservable cognitive mechanisms: The contribution of mathematical models. In *Neuroimaging of human memory: Linking cognitive processes*

to neural systems (pp. 25–41). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199217298.003.0003

van Dijck, J.-P., Abrahamse, E. L., Acar, F., Ketels, B., & Fias, W. (2014). A working memory account of the interaction between numbers and spatial attention. *Quarterly Journal of Experimental Psychology*, *67*(8), 1500–1513. https://doi.org/10.1080/17470218.2014.903984

van Dijck, J.-P., Abrahamse, E. L., Majerus, S., & Fias, W. (2013). Spatial attention interacts with serial-order retrieval from verbal working memory. *Psychological Science*, *24*(9), 1854–1859. https://doi.org/10.1177/0956797613479610

van Dijck, J.-P., & Fias, W. (2011). A working memory account for spatial-numerical associations. *Cognition*, *119*(1), 114–119. https://doi.org/10.1016/j.cognition.2010.12.013

van Dijck, J.-P., Gevers, W., & Fias, W. (2009). Numbers are associated with different types of spatial information depending on the task. *Cognition*, *113*(2), 248–253. https://doi.org/10.1016/j.cognition.2009.08.005

van Dijck, J.-P., Ginsburg, V., Girelli, L., & Gevers, W. (2015). Linking numbers to space: From the mental number line towards a hybrid account. In R. C. Kadosh & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 89–105). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.020

Van Opstal, F., Fias, W., Peigneux, P., & Verguts, T. (2009). The neural representation of extensively trained ordered sequences. *NeuroImage*, *47*(1), 367–375. https://doi.org/10.1016/j.neuroimage.2009.04.035

Verguts, T., Fias, W., & Stevens, M. (2005). A model of exact small-number representation. *Psychonomic Bulletin & Review*, *12*(1), 66–80. https://doi.org/10.3758/bf03196349

Viarouge, A., Hubbard, E. M., & McCandliss, B. D. (2014). The cognitive mechanisms of the SNARC effect: An individual differences approach. *PLOS ONE*, *9*(4), Article e95756. https://doi.org/10.1371/journal.pone.0095756

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043

Vicovaro, M., Boscariol, R., & Dalmaso, M. (2025). A SNARC-like effect for visual speed. *Attention, Perception, & Psychophysics*, *87*(3), 1042–1059. https://doi.org/10.3758/s13414-025-03012-x

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. https://doi.org/10.1016/j.tics.2003.09.002

Wang, Q., Ma, B., Li, M., Ye, X., & Yang, J. (2024). Encoding numbers when numbers rotate to the left or right space: Evidence from the Spatial-Numerical Association of Response Codes (SNARC) effect. *Studia Psychologica*, *66*(4), 267–285. https://doi.org/10.31577/sp.2024.04.905

Wasner, M., Moeller, K., Fischer, M. H., & Nuerk, H.-C. (2015). Related but not the same: Ordinality, cardinality and 1-to-1 correspondence in finger-based numerical representations. *Journal of Cognitive Psychology*, *27*(4), 426–441. https://doi.org/10.1080/20445911.2014.964719

Wilson, R. E., Thompson, R. J., & Vazire, S. (2017). Are fluctuations in personality states more than fluctuations in affect? *Journal of Research in Personality*, *69*, 110–123. https://doi.org/10.1016/j.jrp.2016.06.006

Wood, G., Willmes, K., Nuerk, H.-C., & Fischer, M. H. (2008). On the cognitive link between space and number: A meta-analysis of the SNARC effect. *Psychology Science*, *50*(4), 489–525.

Wühr, P., & Richter, M. (2022). Relative, not absolute, stimulus size is responsible for a correspondence effect between physical stimulus size and left/right responses. *Attention, Perception, & Psychophysics*, *84*(4), 1342–1358. https://doi.org/10.3758/s13414-022-02490-7

Wühr, P., & Seegelke, C. (2018). Compatibility between physical stimulus size and left-right responses: Small is left and large is right. *Journal of Cognition*, *1*(1), 17. https://doi.org/10.5334/joc.19

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193. https://doi.org/10.1016/0010-0277(90)90003-3

Yılmaz, H. (2011). Learning to read (again): The social experiences of Turkey's 1928 alphabet reform. *International Journal of Middle East Studies*, *43*(4), 677–697. https://doi.org/10.1017/S0020743811000900

Zayas, V., Sridharan, V., Lee, R. T., & Shoda, Y. (2019). Addressing two blind spots of commonly used experimental designs: The highly-repeated within-person approach. *Social and Personality Psychology Compass*, *13*(9), 12487. https://doi.org/10.1111/spc3.12487

Zebian, S. (2005). Linkages between number concepts, spatial thinking, and directionality of writing: The SNARC effect and the REVERSE SNARC effect in English and Arabic monoliterates, biliterates, and illiterate Arabic speakers. *Journal of Cognition and Culture*, *5*(1–2), 165–190. https://doi.org/10.1163/1568537054068660

Zhang, Z., Chen, Q., Wang, C., Zhao, D., & Pan, Y. (2025). Decoding the temporal dynamics of numerical semantics: Early cardinal activation and the critical role of ordinal processing in spatial-numerical associations. *Cognition*, *264*, Article 106232. https://doi.org/10.1016/j.cognition.2025.106232

Zohar-Shai, B., Tzelgov, J., Karni, A., & Rubinsten, O. (2017). It does exist! A left-to-right spatial–numerical association of response codes (SNARC) effect among native Hebrew speakers. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(4), 719–728. https://doi.org/10.1037/xhp0000336

Zorowitz, S., & Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *8*(8), 789–797. https://doi.org/10.1016/j.bpsc.2023.02.004

Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel „Resolving controversies about Spatial-Numerical Associations: Insights from different tasks, cultures, stimuli, and time points“ selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen (alternativ: Zitate) als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Tübingen, 02.10.2025

_____

Lilly Zoe Roth

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my three supervisors, Prof. Dr. Hans-Christoph Nuerk, Prof. Dr. Ulf-Dietrich Reips, and Assoc. Prof. Dr. Krzysztof Cipora, for their invaluable guidance, encouragement, and critical insights. I am thankful for everything I have learned from each of them.

I am also very grateful to my dear coworkers at Uni Tübingen, especially for all the scientific and non-scientific chats during countless shared lunch breaks. I extend this gratitude to my international collaborators and the MCLS community for sharing their inspiring perspectives, allowing me to collaborate with them, and inviting me on lab visits. In particular, I am grateful to Prof. Dr. Susan C. Levine, who generously hosted me for an extended stay at UChicago.

Next, I would like to thank the two “SNARC besties,” Assoc. Prof. Dr. Krzysztof Cipora and Prof. Dr. Jean-Philippe van Dijck, as well as my colleague Dr. Elena Albu and my close friend Mara Seyfert for providing content-related feedback on different parts of this dissertation. I would also like to thank Prof. Dr. Martin V. Butz and Ghada Hassan for introducing me to the world of cognitive modeling and helping me contextualize the SNARC and SNARC-like effects. Moreover, I would like to thank Nina Köhler for proofreading the English, and Prof. Dr. Hans-Christoph Nuerk, Prof. Dr. Ulf-Dietrich Reips, and Assoc. Prof. Dr. Michele Vicovaro for reviewing this dissertation.

I also gratefully acknowledge the financial support of this work by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through the project “Replicability of Fundamental Results on Spatial-Numerical Associations in Highly Powered Online Experiments (e-SNARC)” (NU 265/8-1) granted to Prof. Dr. Hans-Christoph Nuerk and Prof. Dr. Ulf-Dietrich Reips in cooperation with Assoc. Prof. Dr. Krzysztof Cipora, and through several projects within the Research Unit FOR2718 “Modal and Amodal Cognition” granted to Prof. Dr. Hans-Christoph Nuerk (NU 265/5-1 and NU 265/5-2) and to Prof. Dr. Martin V. Butz (BU 1335/12-1 and BU 1335/12-2). I additionally thank the UKRI Economic and Social Research Council for financially supporting Assoc. Prof. Dr. Krzysztof Cipora (ES/W002914/1).

My deepest thanks go to my family: Mama, Papa, and Clara, for their unconditional love and support, and for always being there. Oma, Holger, and Opa, for their encouragement and pride in me. Nici and Erich, for all the inspiring discussions during our mountain hikes.

Finally, a very special thanks goes to my favorite people, Lulu and Philipp, for their invaluable humor, loyalty, honesty, and all the love they give me every single day. I also want to thank the Albi-Marktplatz-Fusion, the Stammi, the Tanztheater Tübingen ensemble, the TV Derendingen volleyball group, the kids' dance group at SSC Tübingen, and all my other friends for their emotional support and the joyful distractions during this journey.

Appendix

A1. Study 1

A1.1. Publication of Study 1

Citation:
Roth, L., Caffier, J., Reips, U.-D., Cipora, K., Braun, L., & Nuerk, H.-C. (2025). True colors SNARC: Semantic number processing is highly automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 51*(12), 1875–1908. https://doi.org/10.1037/xlm0001431

Copyright:
© 2025 American Psychological Association
Used for this dissertation with permission of the American Psychological Association; reprinted from “True colors SNARC: Semantic number processing is highly automatic” by Roth et al., Journal of Experimental Psychology: Learning, Memory, and Cognition, volume 51, issue 12, 2025; permission conveyed through Copyright Clearance Center, Inc.

Journal of Experimental Psychology: Learning, Memory, and Cognition

True Colors SNARC: Semantic Number Processing Is Highly Automatic

Lilly Roth, John Caffier, Ulf-Dietrich Reips, Krzysztof Cipora, Lydia Braun, and Hans-Christoph Nuerk

Online First Publication, February 20, 2025. https://dx.doi.org/10.1037/xlm0001431

CITATION

Roth, L., Caffier, J., Reips, U.-D., Cipora, K., Braun, L., & Nuerk, H.-C. (2025). True colors SNARC: Semantic number processing is highly automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://dx.doi.org/10.1037/xlm0001431

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: Learning, Memory, and Cognition

© 2025 American Psychological Association
ISSN: 0278-7393

https://doi.org/10.1037/xlm0001431

True Colors SNARC: Semantic Number Processing Is Highly Automatic

Lilly Roth[1], John Caffier[2], Ulf-Dietrich Reips[2], Krzysztof Cipora[3], Lydia Braun[1], and Hans-Christoph Nuerk[1, 4, 5]

[1] Department of Psychology, University of Tübingen

[2] Department of Psychology, University of Konstanz

[3] Department of Mathematics Education, Centre for Mathematical Cognition, Loughborough University

[4] Learning, Educational Achievement, and Life Course Development Graduate School and Research Network, University of Tübingen

[5] German Center for Mental Health, Berlin, Germany

Numbers are highly relevant in our everyday lives. Besides intentionally processing number magnitude when necessary, we often automatically process it even when not required. The SNARC (Spatial–Numerical Association of Response Codes; Dehaene et al., 1993) effect, describing faster left-/right-sided responses to smaller/larger numbers, respectively, provides evidence for this automaticity. It arises in semantic number-processing tasks both, when number magnitude is task-relevant (e.g., magnitude classification) and task-irrelevant (e.g., parity judgment). However, findings on the SNARC effect in tasks requiring the processing of nonsemantic number features are mixed: While it has been observed in orientation judgment tasks, it was mostly absent in color judgment tasks. Importantly, previous studies were underpowered or did not control for confounding variables. In two highly powered online experiments, we found a small but significant SNARC effect in both nominal color judgment (cyan vs. yellow; slope = −1.71 ms) and color intensity judgment (light cyan vs. dark cyan; slope = −1.13 ms) of Arabic digits from 1 to 9 excluding 5, which did not significantly differ in size. Further, we found little evidence for the Linguistic Markedness of Response Codes (i.e., faster left-/right-sided responses to odd/even numbers, respectively; Nuerk et al., 2004) effect. Moreover, the odd effect (i.e., faster responses to even than to odd numbers; Hines, 1990) was detected. Taken together, both magnitude and parity are processed automatically even if participants respond to physical nonsemantic and nonspatial number features, but the spatial mapping seems more automatic for magnitude than for parity.

Keywords: automaticity, Spatial–Numerical Associations, Spatial–Numerical Association of Response Codes effect, Linguistic Markedness of Response Codes effect, color judgment

Supplemental materials: https://doi.org/10.1037/xlm0001431.supp

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Miriam Rosenberg-Lee served as action editor.

Lilly Roth https://orcid.org/0000-0002-2690-0504
John Caffier https://orcid.org/0000-0002-0475-9167
Ulf-Dietrich Reips https://orcid.org/0000-0002-1566-4745
Krzysztof Cipora https://orcid.org/0000-0003-0077-9336
Lydia Braun https://orcid.org/0009-0007-2073-5521
Hans-Christoph Nuerk https://orcid.org/0000-0002-0331-7498

All the authors take responsibility for the integrity of the data and the accuracy of the data analysis. The authors declare no financial or nonfinancial conflicts of interest with the content of this article.

This research was funded by the Deutsche Forschungsgemeinschaft project "Replicability of Fundamental Results on Spatial–Numerical Associations in Highly Powered Online Experiments (e-SNARC)" (NU 265/8-1 and RE 2655/3-1) granted to Hans-Christoph Nuerk and Ulf-Dietrich Reips supporting Lilly Roth as well as John Caffier, with the assistance of Krzysztof Cipora as a cooperation partner. Krzysztof Cipora is supported by the UKRI Economic and Social Research Council (Grant number ES/W002914/1). The authors thank Sebastian Sandbrink for English proofreading of the article. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Conceptualization: Krzysztof Cipora, Hans-Christoph Nuerk, and Ulf-Dietrich Reips; data curation: Lydia Braun, Krzysztof Cipora, Hans-Christoph Nuerk, Ulf-Dietrich Reips, and Lilly Roth; formal analysis: Lydia Braun and Lilly Roth; funding acquisition: Krzysztof Cipora and Hans-Christoph Nuerk, Ulf-Dietrich Reips; investigation: Lydia Braun, John Caffier, and Lilly Roth; methodology: Krzysztof Cipora, Hans-Christoph Nuerk, Ulf-Dietrich Reips, and Lilly Roth; project administration: Hans-Christoph Nuerk, Ulf-Dietrich Reips, and Lilly Roth; resources: Hans-Christoph Nuerk and Ulf-Dietrich Reips; software: John Caffier and Ulf-Dietrich Reips; supervision: Krzysztof Cipora, Hans-Christoph Nuerk, and Ulf-Dietrich Reips; validation: Krzysztof Cipora, Hans-Christoph Nuerk, Ulf-Dietrich Reips, and Lilly Roth; visualization: Lydia Braun and Lilly Roth; writing–original draft: Lilly Roth; writing–

continued

Information can be processed either automatically or intentionally. According to the instance theory for automatic cognitive processing proposed by Logan (1988), automaticity relies on separate representations (i.e., *instances*) of a task. Specifically, processing can become automatic when instances stored in long-term memory are repeatedly retrieved, and the encoding of the processed elements is practiced. Cognitive processing is faster and less error prone, if it is automatic than if it is intentional. Automaticity is therefore especially important for information that is relevant in many contexts in everyday life, such as numerical information.

Automatic Number Processing

Humans deal with numbers in many situations in everyday life. On the one hand, an external symbolic numeration system (e.g., the Arabic system) is crucial for us to represent numbers, and on the other hand, an internal mental representation (e.g., a mental number line) is necessary for us to understand the meaning of quantities. Western cultures typically use the Arabic system as an external symbolic numeration system, which is based on digits and on the place-value system. The 10 digits from 0 to 9 express quantities, and the position of each digit within a number represents their power with a base 10 (Tzelgov et al., 2015). However, this external system does not correspond to the internal mental representation of numbers, which is claimed to be spatial. More precisely, numerical cognition researchers speak of a mental number line which aligns numbers on a horizontal continuum from left (small numbers) to right (large numbers), at least in Western cultures (as originally proposed by Restle, 1970). Regarding the automaticity of number processing in the Arabic number system, mental representations of separate symbolic numbers can be understood as *instances* (as defined by Logan, 1988).

Tzelgov et al. (2015) distinguish between numbers whose meaning is stored in long-term memory, which they call *primitives* of the numeration system, and numbers whose meaning needs to be generated when required, which they call *nonprimitives*. Single-digit numbers in the Arabic system can be considered as such primitives, because they can be directly mapped onto the mental number line. This straightforward link between the external symbolic numeration system and the internal mental representation of single-digit numbers is in line with the instance theory: Single-digit numbers are primitives and therefore processed automatically, because their meaning is stored in long-term memory. Importantly, automatic processing happens even when irrelevant to the task at hand (Tzelgov et al., 1992). Regarding the automaticity of number magnitude processing, different claims have been made in the literature. On the one hand, early studies by Dehaene and colleagues have demonstrated automatic processing of number magnitude through neural adaptation and priming (Dehaene & Akhavein, 1995; Dehaene et al., 1998; Naccache & Dehaene, 2001a, 2001b). On the other hand, Cohen (2009) stated that number magnitude is not automatically activated in every context and that even single-digit numbers can be processed nonsemantically. To complete their task, participants could either process the magnitude or the physical shape of single-digit numbers presented on the screen; results suggested that responses were determined by physical shape rather than magnitude. A class of phenomena that clearly reflect automatic number magnitude processing are Spatial–Numerical Associations (M. H. Fischer & Shaki, 2014; Toomarian & Hubbard, 2018). They denote systematic associations of numerical information such as cardinality, ordinality, or parity with space (Cipora et al., 2018; Cipora, He, et al., 2020; Patro et al., 2014). The hallmark directional Spatial–Numerical Association is the Spatial–Numerical Association of Response Codes (SNARC) effect, which denotes that in left-to-right reading cultures, participants respond faster to small/large number magnitudes on the left/right side, respectively (Dehaene et al., 1993). It is typically quantified by subtracting the average response time (RT) with the left hand from the average RT with the right hand for each number separately (Fias et al., 1996). Decreasing average differences in response times between hands (dRT) with increasing number magnitude reflect the typical SNARC effect. The SNARC effect reflects automatic number processing, because the mental spatial mapping of number magnitude arises even when number magnitude is not required for the task to be solved. Note that automatic number processing is not the only prerequisite for the SNARC effect to arise: Spatial-directional processing must also be automatic, when space is task-irrelevant (Cipora, He, et al., 2020). This is important when discussing the automaticity of the SNARC, because most previous work used either semantic numerical features (and automatic activation of space or other semantic number representations) or spatial features (and automatic activation of numbers or other spatial codes) as task-relevant features.

In an interim summary, we can conclude that the automaticity of number processing is still somewhat controversial and may depend on task and context. This had led some scholars to recently propose that "processing the numerical magnitude is partly automatic" (Avitan et al., 2024, p. 1).

Automatic Number Processing as Reflected by the SNARC Effect

Typically, the SNARC effect is investigated in studies with judgments on semantic properties of the presented numerical stimuli (see Wood et al., 2008, for a meta-analysis). Interestingly, apart from judgments on number magnitude itself (e.g., classification of numbers as smaller or larger than five), the spatial mapping of number magnitude also arises automatically in judgments on other semantic properties of the presented numerical stimuli. For example, in a parity judgment task, Arabic single-digit numbers are presented on the screen, and participants are asked to judge whether these numbers are odd or even, using a binary manual response–key setup (Dehaene et al., 1993). The SNARC effect is highly replicable

review and editing: Lydia Braun, Krzysztof Cipora, Hans-Christoph Nuerk, and Ulf-Dietrich Reips.

The data are available at https://osf.io/xpyf7/; https://osf.io/jv4as/; https://osf.io/mwqh8/

The experimental materials are available at https://exp.wextor.eu/sna/numcog1a-english/; https://exp.wextor.eu/sna/numcog1b-english/

The preregistered design is available at https://doi.org/10.17605/OSF.IO/F2GB8; https://doi.org/10.17605/OSF.IO/VBA7N

Correspondence concerning this article should be addressed to Hans-Christoph Nuerk, Department of Psychology, University of Tübingen, Schleichstraße 4, 72076 Tübingen, Germany. Email: hc.nuerk@uni-tuebingen.de

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

in this task, although magnitude is irrelevant for the task execution, and therefore demonstrates an "automatic activation of number magnitude" (Dehaene & Akhavein, 1995, p. 317).

Importantly, the SNARC effect has not only been shown in tasks that require semantic number processing, where the identity and certain attributes of numbers must be identified and processed as such in parity judgment. The effect has also been detected in tasks that require only the processing of nonsemantic number features; in these experiments, the identity of the number and all its numerical (semantic) attributes are irrelevant for the task. For instance, the SNARC effect arises, when participants are asked to judge the occurrence of a letter in the names of visually presented numbers (Fias et al., 1996; Experiments 2 and 3) or physical visual features of numbers themselves or of superimposed objects (examples discussed in detail below). In these tasks, not only number magnitude but also all other semantic features of the presented numbers are entirely task-irrelevant. Thus, finding the SNARC effect in such tasks indicates an even higher degree of its automaticity than finding it in tasks requiring semantic number processing. Hence, using tasks with judgments of nonsemantic number features permits investigating the robustness of automatic number processing.

Crucially, evidence for the SNARC effect in tasks that require only the processing of nonsemantic features of numerical stimuli in previous studies is mixed. The SNARC effect has consistently been found in judgments of the orientation of Arabic digits themselves or of superimposed objects. Fias et al. (2001) detected it when participants were to judge whether triangles superimposed on the numbers were pointing upward or downward (Experiment 1) or whether lines superimposed on the numbers were horizontal or vertical (Experiment 4). Moreover, Lammertyn et al. (2002) found a SNARC effect when participants were to judge whether numbers were presented upright or tilted 10° (orientation condition); importantly, in this case, at least the numerical form must be processed before determining whether this form is upright or tilted. In terms of the triple-code model (according to which numbers are mentally represented in an Arabic, verbal, or analogical code; Dehaene, 1992), at least one code of number representation is still processed, even if it is not a numerical feature itself like magnitude and parity. What is more, these tasks require some sort of spatial processing (e.g., orientation), which, according to the neural overlap hypothesis (Fias et al., 2001; details see below), could activate spatial processing. Thus, while semantic of numbers is not processed in these tasks, spatial processing (as well as processing of number form) is still required. To demonstrate full automaticity of space–number associations, it would therefore be preferable to use tasks or paradigms that require neither semantic number processing nor spatial processing. Color judgment tasks are such paradigms.

No Automatic Number Processing in Color Judgment Tasks?

As color judgment tasks require neither semantic number processing nor spatial processing, neither part of the Spatial–Numerical Associations indexed by the SNARC effect can be triggered by such tasks. If the SNARC effect was fully automatic, it should be observed even in such nonsemantic and nonspatial tasks involving numbers. In the following, we review previous experiments investigating the SNARC effect in color judgment. Table 1 provides an overview of these previous experiments.

In the first such investigation, Fias et al. (2001) and Lammertyn et al. (2002) did not detect the SNARC effect in color judgment tasks, although they consistently found it in orientation judgment tasks. Fias et al. investigated it first in nominal color judgment, where participants decided whether Arabic digits were red or green (Experiment 2), and second in color intensity judgment, where participants decided whether Arabic digits were light cyan or dark cyan (Experiment 3). Lammertyn et al. (2002) replicated the color intensity judgment (in their color condition), while using the exact same upright or tilted stimuli as in their other task (in their orientation condition) to avoid confounds, and could also not find a SNARC effect. Moreover, Fias et al. did not find a SNARC effect when participants judged whether the shape in the background of Arabic digits was a circle or a square (Experiment 5). As a possible explanation of their findings, Fias et al. formulated the *neural overlap hypothesis*: They argued that the SNARC effect is only present, if there is an overlap of neural structures for processing relevant and irrelevant information. Number processing and orientation processing both involve the parietal cortex (dorsal pathway), while processing of color and shape does not (ventral pathway). Instead of the similarity of stimulus and response, the authors suggested that the similarity of involved neural structures are decisive for the presence or absence of interference reflected in the SNARC effect. However, Keus (2006) and Toomarian and Hubbard (2018) offered arguments against the neural overlap hypothesis: First, there might in fact be neural overlap between number and color processing, because color processing also seems to involve the parietal cortex (Claeys et al., 2004), and second, the SNARC might be related to response selection rather than to stimulus features (Keus et al., 2005; Keus & Schwarz, 2005). Furthermore, number form did not need to be processed in most experiments obtaining null results.

It would be highly relevant for the field of numerical cognition, if the SNARC effect was only elicited in some but not all judgments of physical nonsemantic number features, because this would mean that the automaticity of number-magnitude processing and of its mental mapping onto space depend on what humans focus on in each situation. Before drawing such impactful conclusions, it is important to make sure that the observations made in Fias et al.'s and Lammertyn et al.'s (2002) studies are replicable. Running power analyses before data collection was not common practice, yet when these studies were conducted. Importantly, absence of an effect should not be interpreted as evidence of no effect, even in highly powered studies. In fact, conclusions drawn from null results obtained in frequentist analyses might be one of several causes of the replication crisis (Maxwell et al., 2015). At the same time, several follow-up studies offered hints toward some form of the SNARC effect in the color judgment tasks.

Nominal Color Judgment (Distinguishing Between Color Hues)

Following Fias et al.'s (2001) null results, further studies have found mixed evidence regarding the SNARC effect in nominal color judgment (see Table 1). Importantly, the studies differed in methodological and design-related aspects.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 1

Previous Experiments Investigating the SNARC Effect in Color Judgment Tasks, Used Sample Sizes, Observed SNARC Effects, and Effect Sizes

Experiment	Task	Sample size	SNARC effect
Cleland and Bull (2019), Experiment 1	Nominal color judgment (blue vs. green) of numbers and symbols	40 adults	$b = -0.013$, $d = 0.035$, $p = .824$
Cleland and Bull (2019), Experiment 5, condition with 0 ms delay in color onset	Nominal color judgment (blue vs. green)	84 adults	$b = -0.068$, $d = 0.154$, $p = .162$
Didino et al. (2019)	Nominal color judgment (red vs. green) with each number always presented in the same color	32 adults	$b = -0.64$, $d = 0.17$, $p = .273$
Fias et al. (2001), Experiment 2	Nominal color judgment (red vs. green)	24 adults	$b = 0.5$, $d = 0.08$ (calculated based on $SD = 6$ reported in the article), $p < .3$
Fias et al. (2001), Experiment 3	Color intensity judgment (light cyan vs. dark cyan)	24 adults	$b = 1.95$, $d = 0.29$ (calculated based on $SD = 6.8$ reported in the article), $p < .1$
Keus and Schwarz (2005), Experiment 3	Nominal color judgment (red vs. green) with lateralized stimulus presentation	16 adults	$b = -1.03$, $d = 0.41$ (calculated based on $SD = 2.5$ reported in the article), only significant in trials compatible with the Simon effect (Simon, 1990; no p value provided)
Lammertyn et al. (2002)	Color intensity judgment (light cyan vs. dark cyan)	20 adults	$b = -0.13$, $d = -0.031$, calculated based on $t(19) = -0.137$ reported in the article, $p = .892$
Hoffmann et al. (2013)	Nominal color judgment (red vs. green) with 200 ms delay in color onset	84 children (5 years old)	$b = -11.80$, $d = 0.39$, calculated based on $t(80) = -3.50$ reported in the article, $p < .001$
Schroeder et al. (2017a), baseline performance without transcranial direct current stimulation	Nominal color judgment (blue vs. green) with numbers 1, 2, 4, 5	two groups of 24 adults	Anodal group: $b = -3.782$, $d = 0.19$ (calculated based on b, n, and p reported in the article for a two-sided t test), $p = .350$, Cathodal group: $b = -2.562$, $d = 0.22$ (calculated based on b, n, and p reported in the article for a two-sided t test), $p = .295$
Schroeder et al. (2017b)	Nominal color judgment (yellow vs. blue) with numbers 1, 2, 4, 5	60 adults	$b = -14.4$, $d = 0.23$, calculated based on $t(59) = 1.76$ reported in the article, $p < .083$

Note. Upper part: experiments run with adults similar to the two experiments in the present study; lower part: experiments run with children or stimulus sets different to Arabic digits from 1 to 9 excluding 5. SNARC = Spatial–Numerical Association of Response Codes.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

First, Keus and Schwarz (2005) found the SNARC effect in nominal color judgment, where stimulus presentation was lateralized instead of central. Note that the response-to-key assignment (green-left and red-right vs. red-left and green-right) was not reversed within but only between participants, just as in Fias et al. (2001). Hoffmann et al. (2013) found the SNARC effect in nominal color judgment conducted with 5-year-old children, but only in children who had previously completed number-magnitude classification. As outlined by Cleland and Bull (2019), this finding is surprising, because the SNARC effect is typically found only at later ages, and the same children did indeed not show the SNARC effect in number-magnitude classification. Importantly, the numbers were first presented in black and switched to red or green only after 200 ms. Hence, number-magnitude classification might have primed processing of number magnitude, and the delayed color onset in nominal color judgment, which might have left enough time to process the number magnitude before judging its font color. Cleland and Bull could not find a SNARC effect in typical nominal color judgment with delayed color onset (Experiments 1 and 5). However, they found the SNARC effect in a go/no-go setup (Experiment 2) with Arabic digits and Greek letters, where color judgment was to be done only on numbers. Moreover, they also found the SNARC effect, when numbers were first presented in black, and their color changed with a delay (trials with a color onset after 0, 200, and 400 ms were intermixed in Experiment 3 and blocked in Experiment 5). Similarly, Lengyel and Krajcsi (2024) found a SNARC effect with a 400-ms color onset delay for blue versus red decisions in symbolic, but not nonsymbolic numbers. These findings suggest that the SNARC effect arises in nominal color judgment only if numbers must be identified as such or if the processing of number magnitude is possible earlier than the processing of color. This is in line with Hoffmann et al.'s (2013) results, suggesting that the SNARC effect only emerges in color judgment with previously enforced number-magnitude processing. Importantly, this prerequisite might be shared with the attentional SNARC effect (Cipora & Nuerk, 2023). To detect true automaticity without such a confound, one therefore needs to avoid enforcing magnitude processing in any task or within the trial before color is judged.

In Schroeder et al.'s (2017a, 2017b) and Didino et al.'s (2019) studies, no significant SNARC effect was detected in nominal color judgment, but results tended to go in the typical direction. Note, however, that Schroeder and colleagues used a smaller than typical number set comprised low-magnitude numbers (1, 2, 4, and 5) and Didino et al. (2019) did not display every number in each of the two colors.

Importantly, the studies summarized above have at least two limitations. First, only small samples were tested (see Table 1). As discussed in the Method section in more detail, at least 265 participants need to be tested to achieve high statistical power for finding a small SNARC effect size. Hence, truly existing effects might have remained undetected due to lack of power in previous studies. Second, in most of the studies, discrimination was to be performed between red and green numbers (see Table 1). This can lead to confounds due to the valence of red and green (association between green, positive, and right, and association between red, negative, and left) and due to the everyday meaning of red and green (e.g., in traffic lights). Moreover, color blindness was not assessed in most studies, and red-green color blindness is the most common color vision defect (Wong, 2011). The average SNARC slope (weighted with sample sizes) for all experiments run with adults and the typical stimulus set from 1 to 9 excluding 5 (listed in the upper part of Table 1) is 0.05 ms.

In summary, the degree of automaticity of Spatial–Numerical Associations in different situations remains unclear; especially in tasks that require neither semantic number processing nor spatial processing. Thus, the question whether space–number processing is fully automatic when neither of its aspects is task-relevant cannot be answered. Ultimately, we have outlined that previous experiments either suffer from low power or design-related confounds. Therefore, a replication of previous color judgment tasks with methodological improvements and high statistical power for small effect sizes is necessary. With the advent of online experimentation, the power issue can be solved easily (Reips, 2009, 2021). Hence, in Experiment 1 of the present study, we asked—after adequate power calculations—340 participants (i.e., many more than in any previous color judgment experiment) to perform nominal color judgment while controlling for confounding variables.

Color Intensity Judgment (Distinguishing Between Luminance Levels)

Apart from nominal color judgments, the SNARC effect can also be investigated in color intensity judgments. Instead of differentiating between two different color hues (e.g., red vs. green), the task is to differentiate between two different luminance levels of the same color hue. As mentioned above, this was done already by Lammertyn et al. (2002), who used light and dark cyan stimuli. The authors did not find a SNARC effect; however, they confounded luminance with the stimuli's orientation.

Color intensity judgment and nominal color judgment differ in one important difference aspect. Namely, Walsh postulated a generalized representation system of magnitude dimensions such as time, space, and quantity in a theory of magnitude (2003), and the luminance of the numbers' font color can also be understood as a magnitude dimension. At the same time, van Dijck et al. (2014) proposed an ordinal position effect along with the working memory account of the SNARC effect, referring to a spatial mapping of items depending on their serial order in working memory (i.e., leftward/rightward shift in mental representation of beginning/late items, respectively). In color intensity judgment (as opposed to nominal color judgment) with a limited amount of different luminances, the stimuli might therefore activate another magnitude or ordinality dimension than numbers (i.e., luminance), which is then potentially mapped onto space as well while luminance is a task-relevant dimension. Crucially, two opposing hypotheses can be derived: This activation could either amplify the activation of number magnitude and therefore trigger or increase its spatial representation as reflected by the SNARC effect or suppress the activation of number magnitude and hence reduce or nullify the SNARC effect.

In Experiment 2 of the present study, we therefore investigated the SNARC effect in color intensity judgment and compared its strength between the two experiments. As for Experiment 1, it was crucial to eliminate confounding variables and achieve high statistical power for small effects while replicating color intensity judgment.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Mental Representation of Luminance

In previous studies, authors have tested predictions derived from a theory of magnitude by investigating interference of different magnitude dimensions (e.g., number magnitude, physical size, conceptual size of word's referents, luminance, space) on the behavioral, developmental, and neural level (e.g., Cleland et al., 2020; Cohen Kadosh et al., 2005, 2008; Pinel et al., 2004; Ren et al., 2011). The term "SNARC effect" has therefore been extended to the "SQUARC (Spatial Quantity Association of Response Codes) effect" (Walsh, 2003). Importantly, it is still under debate whether the emergence of SNARC-like effects (i.e., spatial mapping of other dimensions than number magnitude) support the idea of a generalized magnitude system (Prpic et al., 2021) or might instead be explained by ordinality (Casasanto & Pitt, 2019; Pitt & Casasanto, 2022; van Dijck et al., 2014). Evidence for spatial mappings of magnitude or ordinality dimensions comes from the spatial mapping for task-relevant (Ren et al., 2011) or task-irrelevant (Richter & Wühr, 2022; Wühr & Richter, 2022) nonnumerical magnitude dimensions such as physical size and conceptual size of words' referents, with smaller/larger stimuli being associated with the left/right, respectively. However, the observed spatial mapping of a physical magnitude dimension in color judgment does not say much about the existence of a numerical SNARC effect in color judgments of Arabic digits, because the comparison of physical magnitude seems more automatic than that of number magnitude (Avitan et al., 2024).

Association Between Luminance and Space

Color intensity judgment in a bimanual setup provides the possibility to investigate the horizontal spatial mapping for luminance, for which opposite patterns have been found. On the one hand, Ren et al. (2011) report that high/low luminance was associated with the left/right, respectively, in a luminance comparison task of circles (Experiment 3). On the other hand, both Fumarola et al. (2014) and Wang et al. (2022) report the reversed association when luminance was task-relevant (color intensity comparison of rectangles), and Fumarola et al. (2014) even when luminance was task-irrelevant (nominal color judgment of rectangles), such that low/high luminance was associated with the left/right, respectively. Fumarola et al. (2014) offered the different background luminance as an explanation for the opposed findings (but see Wang et al., who used a light background like Ren et al., 2011, but found the same pattern as Fumarola et al., 2014).

Another difference between studies was that Ren et al. (2011) used achromatic stimuli (different levels of gray), while Fumarola et al. (2014) and Wang et al. (2022) used chromatic stimuli (red and green). In addition to associations being reflected on a behavioral level, recent studies with chromatic stimuli (red and green squares) suggest that associations between luminance and space are also reflected on an affective level, such that participants prefer low/high luminance on the left/right side, respectively (Gusmão et al., 2022; Löffler et al., 2022). In the present study, the association between luminance and space in chromatic stimuli was investigated with a neutral background.

Association Between Luminance and Number Magnitude

Importantly, luminance might not only be associated with space but also with number magnitude. Pinel et al. (2004) found that small/large number magnitude was associated with low/high luminance, respectively, in number-magnitude but not luminance comparison. This further challenges the neural overlap hypothesis by Fias et al. (2001), because Pinel et al. (2004) observed no or only little neural overlap between number magnitude and luminance. Similarly, Cohen Kadosh, Cohen Kadosh, and Henik (2008) found an association between small/large number magnitude with low/high luminance, respectively, both in number-magnitude and luminance comparison. Importantly, Cohen Kadosh and Henik (2006) had previously found the reversed association. In sum, an association between luminance and number magnitude might exist, but its direction remains unclear. We therefore investigated the association in an exploratory analysis.

Automatic Number Processing Reflected by the MARC Effect

Apart from the SNARC effect, the MARC (Linguistic Markedness of Response Codes; Nuerk et al., 2004) effect can also be tested in both color judgment tasks. This effect describes faster left-/right-sided responses to odd/even numbers, respectively. The MARC effect is relatively consistently found in the parity judgment task (Cipora, Soltanlou, et al., 2019; Landy et al., 2008; Nuerk et al., 2004). However, to our knowledge, the MARC effect has not been explored yet in tasks that require only the processing of nonsemantic number features such as in color judgment tasks. Interestingly, a descriptive observation of the dRT pattern in Figure 6 by Didino et al. (2019) points toward the MARC effect (but the dRT patterns of Fias et al., 2001, and Lammertyn et al., 2002, do not). Hence, the MARC effect was tested in both our experiments to investigate whether not only number magnitude but also other number attributes like parity are automatically associated with directional space.

Aims and Hypotheses of the Current Study

To summarize previous studies, the SNARC effect has been demonstrated in some tasks that require processing only nonsemantic number features. Several findings have challenged the neural overlap hypothesis, which is, to our knowledge, the only theory predicting that the SNARC effect is not elicited in color judgments. We therefore did not see any reason for the SNARC effect to be absent in color judgment while being present in judgments of other nonsemantic number features such as orientation. Moreover, evidence for the SNARC effect in color judgment had already been found in some previous studies and might just have remained undetected in others due to confounds caused by the study design or due to lack of power for detecting small effect sizes. Hence, we hypothesized to observe a SNARC effect in both nominal color judgment and color intensity judgment, even if no semantic number processing is required.

The aim of the present study was to find out whether the SNARC effect arises even when only the processing of nonsemantic number features is required. For this, two preregistered online experiments were run: nominal color judgment (Experiment 1) and color intensity judgment (Experiment 2) of Arabic single-digit numbers. In most previous studies using these two tasks, no SNARC effect had been detected. However, it remained unclear whether there is a true null effect or whether a small effect could not be found due to a

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

lack of statistical power, as discussed above. A further limitation of previous studies is that the number of repetitions per experimental cell (Number × Color × Response-to-Key Assignment) was low. In the present study, a sufficiently large sample and a higher number of repetitions per experimental cell ensured high statistical power. On top of null hypothesis testing, we employed Bayesian analysis to differentiate evidence of absence of an effect from absence of evidence for an effect.

Moreover, we adapted the stimuli and background to minimize confounds. First, to avoid effects of valence or everyday life meaning of red and green and problems due to red-green color blindness, yellow and cyan stimuli were used in nominal color judgment. Further, confounds related to luminance of the stimuli and the backgrounds were minimized by manipulating only the color hue in nominal color judgment (Experiment 1), while using the same luminance for both colors and the background. In color intensity judgment (Experiment 2), confounds were minimized by manipulating only the luminance while keeping the color hue of the stimuli constant, and by using the average luminance of the stimuli for the background.

To conclude, the present study aimed at investigating to which degree single-digit numbers (i.e., so-called primitives between 1 and 9, as defined by Tzelgov et al., 2015) are automatically processed. Specifically, our design allowed to test the automaticity of number processing when numbers were semantically not relevant for the task. Hence, the present study allows disentangling previous, sometimes contradictory, claims about automatic human cognition.

Method

The ethics committee of the University of Tübingen's Department of Psychology had approved this study before data collection. We preregistered the hypotheses, method, data preprocessing, and data analysis for both experiments (see Roth et al., 2022a, for nominal color judgment; and Roth et al., 2022b, for color intensity judgment).

Participants

The necessary sample sizes for the two experiments were determined with an a priori power analysis. To ensure a power of $1 - \beta = .90$, to detect an effect of the size Cohen's $d = 0.20$ at a significance level of $\alpha = .05$ for a two-sided one sample t test (to test whether the SNARC slopes differ significantly from zero), the sample needed to consist of at least 265 participants in each experiment. Our aim was to collect at least 300 complete data sets. We did not know how high the dropout rate would be in the online studies (for a review and discussion of dropout in online experiments, see Reips, 2021). To reach the required sample size, we therefore decided to calculate with a rather high dropout rate of 15%. Thus, 360 participants were recruited per experiment via the recruiting platform Prolific, as preregistered. To address a variety of languages using the Latin alphabet with a left-to-right reading and writing direction, both experiments were created in six language versions, namely, English, German, Italian, Polish, Portuguese, and Spanish. We recruited 60 native speakers of the respective language for each of these versions. The expected participation duration was a maximum of 25 min, and participants were compensated with a payment of 4 pound sterling accordingly (approximately 5 euros or 5 U.S. dollars at the time of data collection). Participants received proportional compensation in case of partial participation.

After applying all exclusion criteria as preregistered, data sets from 340 participants were analyzed with a mean age of 28.9 years ($SD = 10.4$) in Experiment 1 and from 312 participants with a mean age of 29.7 years ($SD = 10.9$) in Experiment 2. Further descriptive information about the two samples is displayed in Table 2, and screenshots of the original items can be found in Appendix A (see Figure A1).

Procedure

The English versions of Experiments 1 and 2 are available at https://exp.wextor.eu/sna/numcog1a-english/ and https://exp.wextor.eu/sna/numcog1b-english/. Experimental materials were set up with WEXTOR (https://wextor.eu; Reips & Neuhaus, 2002), a tool using HTML and Javascript designed to assist researchers in creating and hosting online experiments in a guided manner. WEXTOR allows to display stimuli and collect response times both on the server and in participants' web browsers. Different web browsers and operating systems may create some variation in timing of stimuli presentation (Garaizar & Reips, 2019); however, the variation would likely be negligible in the repeated-measures design discussed here. Moreover, the between-subjects variation resulting from participations on different devices under different environmental conditions increases ecological validity and therefore supports generalizability of the results (Reips, 2002). Additionally, screenshots of the experiment pages are provided in Appendices A and B as well as on the Open Science Framework (Roth et al., 2021). On the first page of each experiment, participants were welcomed and presented with information about the study (duration, procedure, voluntariness of participation, possibility to cancel participation at any time, anonymous data collection and storage, contact information of the research team, etc.). Moreover, participants were informed that they needed a computer (desktop or laptop) to take part in the study, and that participation from mobile devices was not possible; both color judgment tasks required a keyboard. Moreover, participants were told that they had to be at least 18 years old to take part and should carry out the study in a quiet environment to avoid distractions. Together with this informed consent, we applied the seriousness check (Reips, 2007, 2009), allowing participants to indicate whether they wanted to participate seriously or simply to view the experiment. This was followed by demographic questions on age, gender, native language, handedness, and finger-counting habits (starting hand and stability).

Next, participants completed a color vision check to ensure they could distinguish the font colors of the numbers. This check consisted of two screens, each containing one row of five digits "5" (Figure B1). In Experiment 1, on each of the two screens, one digit was yellow, and one digit was cyan, while the other three digits were gray (of the same luminance as yellow and cyan). Participants were asked to click once on the yellow and once on the cyan number. In Experiment 2, each of the two screens displayed four times the digit 5: One digit was light cyan, and one digit was dark cyan (colors used in the experimental task), while one of the remaining two digits was light gray and the other was dark gray (of the same luminance as light and dark cyan). Participants were asked to click once on the light cyan and once on the dark cyan number. Note that the word "blue" was used instead of "cyan" in both experiments, because the latter is not frequently used in some languages. In both experiments, participants

Table 2
Descriptive Self-Reported Information About the Samples in Both Experiments

Demographic item	Answer	Experiment 1	Experiment 2
Gender[a]	Woman	134	134
	Man	202	172
	Diverse	4	4
First language[b]	English	62	53
	German	60	50
	Italian	55	49
	Polish	56	55
	Portuguese	59	52
	Spanish	60	59
Handedness	Right-handed	301	273
	Left-handed	28	28
	Ambidextrous	11	11
Finger-counting starting hand[a]	Right-starters	146	157
	Left-starters	133	87
	No preference	60	66
Stability of finger-counting[a]	Always	92	100
	Mostly	152	122
	Slightly more often	36	23
	No preference	60	65
Synesthesia	Based on question[c]	50	43
	Based on comment[c]	30	40
Math skills (0–400)	*M*	247.16	246.00
	SD	85.90	88.08

Note. The final samples consisted of $N_1 = 340$ participants in Experiment 1 and $N_2 = 312$ participants in Experiment 2.
[a] In these demographic items, participants had the option to choose the option "I prefer not to answer." Therefore, the number of all answers is not necessarily equal to the sample size in the respective experiment.
[b] In both experiments, some participants indicated to be bilingual. They participated in one of their two native languages but indicated both in the demographic item. [c] Participants were asked whether some of the numbers between 1 and 9 are generally represented in certain colors in their mind. "Based on question" reflects the number of participants who answered *yes*, while "based on comment" reflects the number of participants who additionally left a comment that points toward synesthesia.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

were familiarized with the stimuli colors in the practice trials. Figure B1 shows a screenshot of the respective screens in the experiments.

After the color vision check, we informed participants that they would need one left and one right response key for the experimental task, and that on their keyboard, they should be located in the same row and about one hand width apart from each other. In online experiments, with participants connecting on their own devices, keyboards may differ dramatically (Reips, 2007). Default keys were D and K (as they are conveniently located on the frequent QWERTY layout), but participants had the option to choose other keys if necessary. This was followed by the instructions for the experimental task. In Experiment 1, participants were to perform nominal color judgment; participants in Experiment 2 were to perform color intensity judgment. The experimental task consisted of two blocks, which were separated by a break of a minimum of 30 s. Both blocks were preceded by a short practice session to accustom participants to the respective response-to-key assignment, which was reversed between blocks. We counterbalanced block order between participants in each language version of each experiment.

In both experiments, after completion of the color judgment task, participants rated their math skills on a Visual Analogue Scale from *very good* to *very bad* in comparison to other people of their age (see Figure A2). A self-report was used instead of a proper math skills assessment, because the variable was only intended for an exploratory analysis (see preregistration) and evidence from previous studies regarding relations between the SNARC effect and math skills is inconclusive (Cipora, He, et al., 2020). Moreover, the study duration was to be kept short, because longer online experiments always elicit the danger of motivational decrease toward the end (Galesic & Bosnjak, 2009). Then, we assessed whether participants were affected by digit-color synesthesia, because this might lead to interference or facilitation in the experimental task (Elias et al., 2003). Hence, participants were asked whether some of the numbers between 1 and 9 are generally represented in certain colors in their minds (see Figure A2). If so, they were asked to further explain in a text field below. Additionally, we assessed data quality with three items. Participants were asked to indicate how noisy the environment was during participation (*silent*, *very quiet*, *fairly quiet*, *fairly noisy*, *very noisy*, or *extremely noisy*), whether there were major distractions during participation (*no major distraction*, *one major distraction*, or *multiple major distractions*), and whether they encountered difficulties during participation (*yes* or *no*, with the option to comment). Lastly, we thanked participants and gave them a completion code for Prolific, so that they could verify their participation.

Several methodological techniques are recommended for online experiments and were implemented (Reips, 2009): To prevent search engine bots (e.g., Googlebot) from submitting data on the experiment, we equipped the HTML experiment pages with a standardized "noindex, nofollow" meta tag. This prompted search engine bots not to index the experiment pages and not to visit

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

subsequent pages. Additionally, we manually checked User Agents and IP addresses but did not find any irregularities, and thus did not have to exclude data sets based on this inspection.

Color Judgment Tasks and Stimuli

Participants in Experiment 1 completed a nominal color judgment task, where they had to decide in each trial whether the number was presented in cyan or yellow, whereas participants in Experiment 2 completed a color intensity judgment task where they had to decide in each trial whether the number was presented in dark cyan or light cyan. Participants were instructed to respond as fast and as accurately as possible. The stimuli were the Arabic numbers 1, 2, 3, 4, 6, 7, 8, and 9, and their magnitude was manipulated in a within-subject design. In each experiment, each number was presented 30 times in each color in a fully randomized order, resulting in 480 experimental trials split into two experimental blocks. In Experiment 1, each number appeared 15 times in each color within the block where participants had to respond with the left key to cyan numbers and with the right key to yellow numbers and 15 times within the block with reversed response-to-key assignment. In Experiment 2, each number appeared 15 times in each color within the block where participants had to respond with the left key to dark cyan numbers and with the right key to light cyan numbers and 15 times in the block with reversed response-to-key assignment. A short practice session with 16 trials preceded each block: Each digit was presented once in each color, the current response-to-key assignment was displayed at the bottom of the screen, and feedback on accuracy was provided.

Each trial started with the presentation of a square (extended ASCII 254, size 72pt) as the eye fixation point for 300 ms. The square was gray in Experiment 1 (RGB: 228, 228, 228) with the same luminance as the cyan and yellow number stimuli (CIELAB: $L = 90.59$) and black in Experiment 2 (RGB: 0, 0, 0). Then the number (Open Sans font, size 72 pt) was presented until the response was given. RTs were measured as the time elapsing from the onset of number presentation until a response key was pressed. In Experiment 1, numbers were presented in cyan (RGB: 0, 255, 255) or in yellow (RGB: 235, 235, 0) against a black (RGB: 0, 0, 0) background. We chose colors with roughly equal luminance for cyan (CIELAB: $L = 90.67$) and yellow (CIELAB: $L = 90.80$). In Experiment 2, numbers were presented in dark cyan (RGB: 0, 88, 93) or in light cyan (RGB: 0, 255, 255) against a gray (RGB: 150, 150, 150) background. A background color with the approximate average luminance (CIELAB: $L = 62.08$) of the dark cyan (CIELAB: $L = 33.26$) and light cyan (CIELAB: $L = 90.67$) stimuli was used. A blank screen in the respective background color followed the stimulus presentation for 500 ms and concluded each trial. Figure B2 illustrates the time course of trials as examples.

Data Preprocessing

Data preprocessing followed the preregistered protocol. In Experiment 1, 385 Prolific users started participating (67 in the English, 64 in the German, 63 in the Italian, 64 in the Polish, 61 in the Portuguese, and 66 in the Spanish version), but 25 of them dropped before or during the nominal color judgment task. Here, we ended up with 360 complete data sets (dropout rate of 6.5%). In Experiment 2, 384 Prolific users started participating (71 in the English, 62 in the German, 61 in the Italian, 63 in the Polish, 62 in the Portuguese, and 65 in the Spanish version), but 21 of them dropped before or during the color intensity judgment task. Here, we ended up with 363 complete data sets (dropout rate of 5.5%).

From the complete data sets, we first excluded data from participants who only wanted to have a look at the experiment and not seriously participate (two in Experiment 1, one in Experiment 2). We preregistered to exclude data from participants who indicate to be less than 18 years old, but none of them did so. Further, we excluded data from participants who described their environment as very or extremely noisy (none in Experiment 1, one in Experiment 2) or who reported multiple major distractions during participation (one in Experiment 1, one in Experiment 2). Additionally, data from participants who failed in the first or second part of the color vision check were excluded (six for cyan and three for yellow in Experiment 1, 32 for dark cyan[1] and seven for light cyan in Experiment 2).

Regarding the remaining data sets (349 for Experiment 1 and 326 for Experiment 2), we first excluded practice trials. RTs were not recorded for 0.4% of the remaining trials in Experiment 1 and for 0.8% of the remaining trials in Experiment 2. Of all experimental trials with recorded RTs, we next excluded incorrectly answered trials from analysis (4.6% in Experiment 1, 5.5% in Experiment 2). Then, trials with reaction times below 200 ms (considered as anticipations; 0.3% of the remaining correctly answered trials in Experiment 1, 0.1% in Experiment 2) and above 1,500 ms (0.8% in Experiment 1, 0.9% in Experiment 2) were excluded from analysis. Next, further outliers were excluded by iterative trimming for each participant separately. RTs that were more than 3 *SD*s above or below the individual mean RT were removed sequentially until no more RTs had to be excluded (2.0% of the remaining trials in Experiment 1, 1.8% in Experiment 2). Finally, we excluded data from participants with less than 75% valid remaining trials (seven in Experiment 1, nine in Experiment 2) or at least one empty experimental cell (Number Magnitude × Response Side; two in Experiment 1, five in Experiment 2). In the end, data sets from 340 participants in Experiment 1 and from 312 participants in Experiment 2 remained for the analysis. Demographic information about these participants can be found in Table 2.

Preregistered Data Analysis

Data analyses were performed with the statistical computing software R (R Core Team, 2022) with the R packages *BayesFactor* (Morey et al., 2022), *clinfun* (Seshan & Whiting, 2023), *dplyr* (Wickham, François, et al., 2023), *GeneNet* (Schaefer et al., 2021), *ggplot2* (Wickham, Chang, et al., 2023), *lmerTest* (Kuznetsova et al., 2017), *plyr* (Wickham, 2022), and *tidyr* (Wickham, Vaughan, et al., 2023). The analysis R scripts as well as the anonymized data are available on the Open Science Framework (Roth et al., 2021).

Repeated-Measures Regression

The key-dependent variable in the following was the mean difference between RTs for the two hands (dRT = RT right hand −

[1] The color vision check in Experiment 2 might have wrongly classified some participants (see Figure B1), because they were instructed to click on the dark "blue" number and might rather have considered the number to be cyan or green instead (one participant addressed this in the comments).

RT left hand). We conducted a repeated-measures regression as described by Lorch and Myers (1990) in order to test the presence of the SNARC and MARC effects on group level. For this, in a first step, dRTs were regressed on number magnitude (Fias et al., 1996) and on number parity (Cipora, Soltanlou, et al., 2019) for each participant separately. A negative number-magnitude slope (referred to as *SNARC slope* in the following) reflects the SNARC effect, because it quantifies an increasing right- over left-hand advantage in milliseconds per increase of number magnitude in one unit. A negative number-parity slope (referred to as *MARC slope* in the following) reflects the MARC effect, because it quantifies the increase in right- over left-hand advantage in milliseconds, when numbers are even instead of odd. The more negative the SNARC or MARC slope is, the stronger the SNARC or MARC effect is. On top of these unstandardized SNARC and MARC slopes, we also used standardized SNARC and MARC slopes obtained by standardizing and Fisher-*z*-transforming the slopes (referred to as *St-SNARC* and *St-MARC* in the results). In contrast to unstandardized slopes, standardized slopes can be interpreted as individual SNARC and MARC "effect sizes" and are unaffected by individual means and individual variances of reaction times (Cipora, van Dijck, et al., 2019; Tzelgov et al., 2013).

To investigate the effects of interest at the group level, we tested the resulting unstandardized and standardized SNARC and MARC slopes from all participants against zero with two-sided one-sample *t* tests in each experiment. Two-sided tests were used, because results from previous literature on color judgment tasks remained inconclusive and the possibility of reversed effects remained. We additionally calculated the Bayes factors (BF_{10}) associated with the Bayesian two-sided one-sample *t* tests, which reflect the probability of the data given the alternative hypothesis in proportion to the probability of the data given the null hypothesis to obtain evidence for both null and alternative hypotheses. For this, we used the default *r* scale of 0.707 as uninformed prior in the Cauchy distribution. We interpreted Bayes factors as proposed by Dienes (2021): A resulting BF_{10} greater than 3 or 10 was treated as moderate respectively strong evidence for the alternative hypothesis compared to the null hypothesis, while a resulting BF_{10} smaller than 1/3 or 1/10 was treated as moderate or strong evidence for the null hypothesis compared to the alternative hypothesis, respectively. A resulting BF_{10} between 1/3 and 3 indicates that the data cannot sensitively distinguish the null hypothesis from the alternative hypothesis, rendering the gathered evidence inconclusive.

H0 Bootstrapping

In addition to group-level SNARC and MARC effects, we investigated participant-level SNARC and MARC effects by means of H0 (i.e., null hypothesis) bootstrapping as proposed by Cipora, van Dijck, et al. (2019) and applied by Roth et al. (2024). This allowed us to find out how many participants showed reliable effects. In the H0 bootstrapping approach, two sets of 30 RTs were randomly sampled with replacement from all valid trials of a given participant for each number separately. The two sets were then treated as left- and right-side responses respectively, and unstandardized SNARC and MARC slopes were calculated. Note that the H0 bootstrapping approach is not suitable for standardized slopes, as discussed by Cipora, van Dijck, et al. (2019). This procedure was repeated 5,000 times to obtain a distribution of SNARC and MARC slopes. We then constructed confidence intervals around 0 containing the mid 90% of these bootstrapped slopes. These 90% represented the range of slopes likely to be observed, if there was no association between magnitude or parity and response side. Observed slopes outside these confidence intervals were considered as reliable (reversed) SNARC or MARC effects. In the analysis, we focus on 90% confidence intervals (as in Cipora, van Dijck, et al., 2019); intervals for other confidence levels are presented in Appendix C (see Table C1 for the SNARC effect and Table C2 for the MARC effect).

Comparison Between Color Judgment Tasks

Lastly, a preregistered two-sided two-sample *t* test with the respective BF_{10} was calculated to compare unstandardized and standardized SNARC and MARC slopes between nominal color judgment and color intensity judgment.

Results

The median duration of complete participations in each of the six language versions in each experiment was between 16 and 20 min. Out of 360 participants in each experiment, default response keys were changed by 26 participants in Experiment 1 and by 33 participants in Experiment 2.

Experiment 1: Nominal Color Judgment

The average RT in all trials remaining after applying exclusion criteria was 443 ms ($SD = 69$ ms). The average RTs after aggregation per participant for each number (1, 2, 3, 4, 6, 7, 8, and 9) were 448, 442, 442, 442, 443, 444, 443, and 443 ms, respectively, with *SD*s lying between 69 and 73 ms per number (see Figure D1).

Repeated-Measures Regression

SNARC Effect. The repeated-measures regression revealed a group-level SNARC effect (see Figure 1).[2] The mean SNARC effect was $\text{slope}_{\text{SNARC}} = -1.71$ ms, $SD = 4.28$, $d = 0.40$, $t(339) = -7.38$, $p < .001$, and the standardized SNARC effect was $\text{slope}_{\text{St-SNARC}} = -0.11$, $SD = 0.24$, $d = 0.45$, $t(339) = -8.33$, $p < .001$, indicating a rather small effect size. $BF_{10} = 4{,}524{,}201{,}011.30$ for SNARC and $BF_{10} = 2{,}468{,}094{,}980{,}829.45$ for St-SNARC indicate strong evidence for the presence of a SNARC effect.

MARC Effect. The repeated-measures regression revealed a group-level MARC effect, with $\text{slope}_{\text{MARC}} = -2.54$ ms, $SD = 20.31$, $d = 0.12$, $t(339) = -2.30$, $p = .022$; and $\text{slope}_{\text{St-MARC}} = -0.04$, $SD = 0.23$, $d = 0.17$, $t(339) = -3.16$, $p = .002$, indicating a left-/right-side advantage for odd/even numbers, respectively. $BF_{10} = 0.83$ for MARC and $BF_{10} = 8.03$ for St-MARC indicate inconclusive evidence and moderate evidence for the effect. This indicates that some participants revealing large standard deviations might drive the unstandardized results with their large reversed MARC effects, while they play a lesser role for the standardized results.

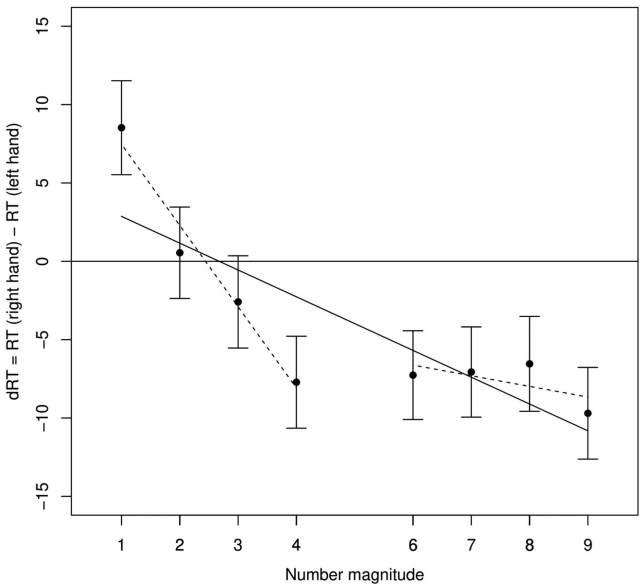
[2] To account for the right-skewed RT distribution, the iterative trimming procedure that sequentially removes all trials outside 3 *SD* from each participant's mean was run with log(RT) instead of raw RT as well. The main data analysis was then rerun, and all four effects (unstandardized and standardized SNARC and MARC) remained significant.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Figure 1

Overall dRTs per Number Magnitude in Nominal Color Judgment (Experiment 1)



Note. Dots represent dRTs averaged across all trials from all participants and reveal potential Spatial–Numerical Association of Response Codes (SNARC) and Linguistic Markedness of Response Codes (MARC) effects. Error bars represent ±1 standard error for the respective number. Solid regression lines represent SNARC slopes for the whole stimulus set, while dashed regression lines represent SNARC slopes calculated separately for the lower range (number magnitudes 1, 2, 3, and 4) and for the higher range (number magnitudes 6, 7, 8, and 9). The MARC effect is reflected by dRTs being on average larger for odd than for even numbers. dRT = differences in response times between both hands; RT = response time.

H0 Bootstrapping

SNARC Effect. 47.94% of the sample revealed a reliable regular SNARC effect, while 17.35% revealed a reliable reversed SNARC effect (positive SNARC slope), and 34.71% indicated no reliable effect (see Figure 2).

MARC Effect. 36.18% of the sample revealed a reliable regular MARC effect, while 24.41% revealed a reliable reversed MARC effect (positive MARC slope), and 39.41% indicated an inconsistent effect (see Figure 3).

Experiment 2: Color Intensity Judgment

The average RT in all trials remaining after applying exclusion criteria was 460 ms (SD = 69 ms), which is slightly higher than in nominal color judgment (Experiment 1). The average RTs after aggregation per participant for each number (1, 2, 3, 4, 6, 7, 8, and 9) were 463, 459, 461, 461, 462, 456, 462, and 462 ms, respectively, with SDs lying between 70 and 76 ms per number (see Figure D2).

Repeated-Measures Regression

SNARC Effect. The repeated-measures regression revealed a group-level SNARC effect (see Figure 4).[3] The mean SNARC effect was $\text{slope}_{\text{SNARC}} = -1.13$ ms, $SD = 3.93$, $d = 0.29$, $t(311) = -5.10$, $p < .001$, and the standardized SNARC effect was $\text{slope}_{\text{St-SNARC}} = -0.08$, $SD = 0.25$, $d = 0.31$, $t(311) = -5.53$, $p < .001$, indicating a rather small effect size. $BF_{10} = 14{,}762.30$ for SNARC and $BF_{10} = 123{,}179.57$ for St-SNARC indicate strong evidence for the SNARC effect.

MARC Effect. In contrast, we did not observe a group-level MARC effect, with $\text{slope}_{\text{MARC}} = 0.24$ ms, $SD = 19.29$, $t(311) = 0.22$, $p = .827$; and $\text{slope}_{\text{St-MARC}} = -0.003$, $SD = 0.23$, $t(311) = -0.22$, $p = .826$. $BF_{10} = 0.07$ for MARC and $BF_{10} = 0.07$ for St-MARC indicate strong evidence against the MARC effect.

H0 Bootstrapping

SNARC Effect. 41% of the sample revealed a reliable regular SNARC effect, while 19% revealed a reliable reversed SNARC effect (positive SNARC slope), and 40% indicated no reliable effect (see Figure 5).

MARC Effect. 28% of the sample revealed a reliable regular MARC effect, while 29% revealed a reliable reversed MARC effect (positive MARC slope), and 44% indicated an inconsistent MARC effect (see Figure 6).

Comparison Between Color Judgment Tasks

To determine whether automatic number processing and the spatial mapping of magnitude and parity are stronger or weaker in color intensity judgment, in which a second magnitude dimension (i.e., luminance) is potentially activated, we compared SNARC and MARC effects between nominal color judgment and color intensity judgment. A significant SNARC effect was present in both experiments. When directly comparing observed SNARC slopes, no significant difference between Experiments 1 and 2 was found, $t(650) = -1.72$, $p = .086$, and the resulting $BF_{10} = 0.37$ reflects inconclusive evidence.

We found a weak but significant group-level MARC effect in Experiment 1, but not in Experiment 2. Importantly, the BF_{10} indicated inconclusive or, at most, only moderate evidence for the MARC effect in Experiment 1. When directly comparing observed MARC slopes, we did not find significant differences between Experiments 1 and 2, $t(650) = -1.76$, $p = .079$, and the resulting $BF_{10} = 0.40$ reflects inconclusive evidence.

Exploratory Analyses

Linear Mixed Models

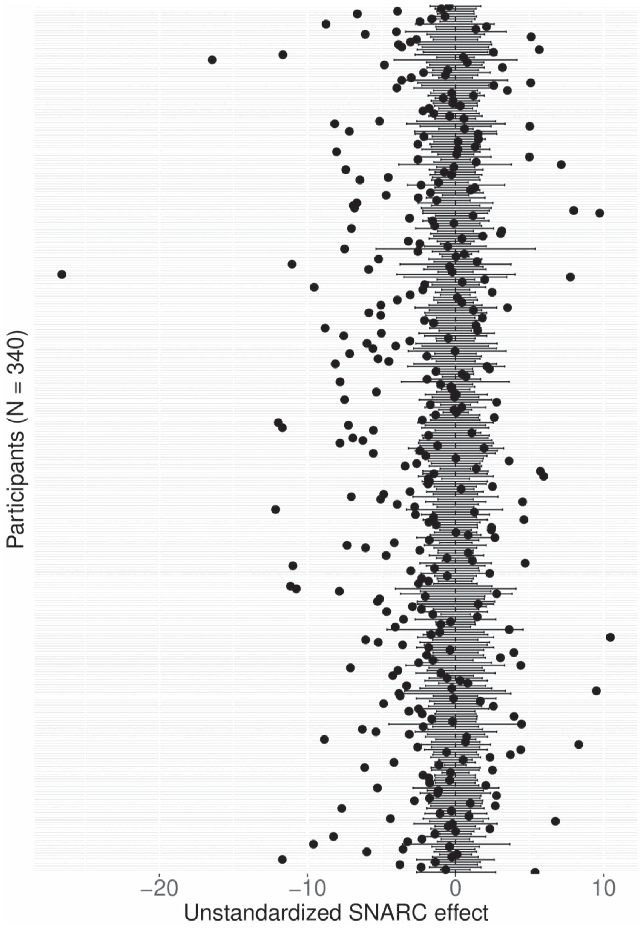
For each experiment, RTs of all participants were regressed on number magnitude, number parity, response side, color, and handedness in a linear mixed model (LMM). Most importantly, we replicated the SNARC effect in both experiments, which was stronger in right-handers than in left-handers. Moreover, an association of even/odd with small/large numbers, respectively (i.e., Parity-Numerical

[3] As in Experiment 1, to account for the right-skewed RT distribution, the iterative trimming procedure was run with log(RT) as well. The main data analysis for Experiment 2 was then rerun, and just as with the iterative trimming based on raw RT, the unstandardized and standardized SNARC effects were significant, whereas the unstandardized and standardized MARC effects were not.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Figure 2

90% H0 Bootstrap Confidence Intervals for the SNARC Slopes in Nominal Color Judgment (Experiment 1)



Note. Confidence intervals for the SNARC effect (*x*-axis) for each participant separately (*y*-axis), with the empirically observed SNARC slope marked as a dot. Overall differences in response times between both hands per number magnitude for participants revealing a reliable regular SNARC effect are illustrated in Figure E1. SNARC = Spatial–Numerical Association of Response Codes; H0 = null hypothesis.

Association of Representational Codes [PNARC] effect; Krajcsi et al., 2018) was found in both experiments (note that Krajcsi et al., 2018, found large interindividual differences regarding the direction of the PNARC effect). All LMM results can be found in Appendix F.

Effects on RTs

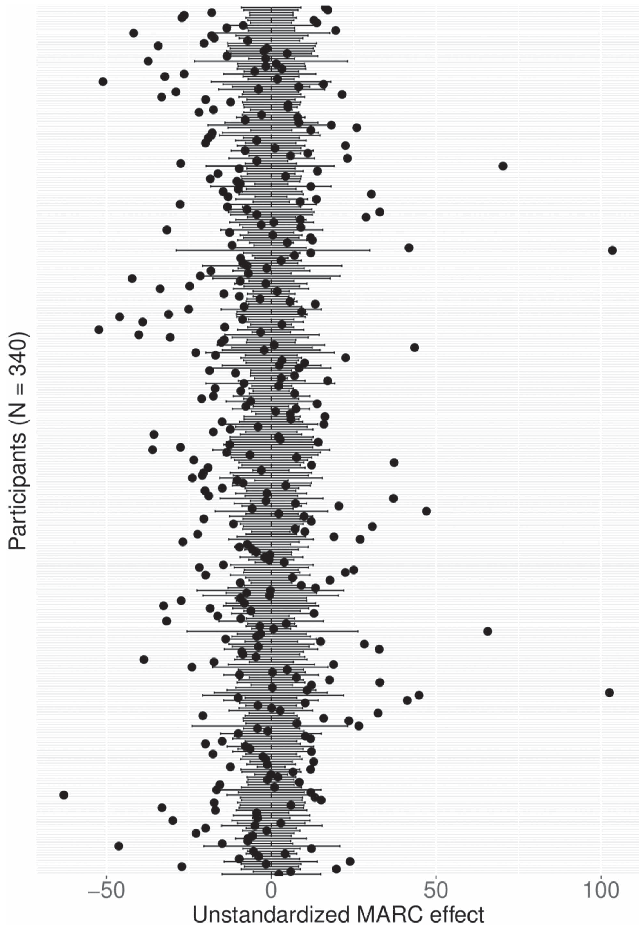
Next, we tested the same effects on RTs as in the LMM above, but this time with the rather traditional analyses, that is, by comparing factor levels via *t* tests or tested regression slopes against 0. The size effect (i.e., slower responses with increasing number magnitude; Fias et al., 2003) was not replicated in any experiment. The odd effect (i.e., faster responses to even than to odd numbers; Hines, 1990) was found in Experiment 1, but not in Experiment 2. In both experiments, RTs were faster with the right than with the left hand, while no evidence was found for them to depend on stimulus color or handedness of the participant. Exact results can be found in Appendix G.

Influences of Participant Characteristics on the SNARC and MARC Effects

No SNARC and MARC effects were detected for participants who seemed to be affected by synesthesia or who were left-handed. However, no evidence was found for differences between these subsamples and participants who did not seem to be affected by synesthesia or who were right-handed. The SNARC and MARC effects did not seem to depend on participants' self-reported math skills or first language. Moreover, there was no difference in the SNARC and MARC effects between participants who indicated to start finger-counting with their right versus left hand and between participants with stable versus unstable habits regarding the starting hand in finger-counting. These results were similar in both experiments and are reported in greater detail in Appendix H.

Figure 3

90% H0 Bootstrap Confidence Intervals for the MARC Slopes in Nominal Color Judgment (Experiment 1)

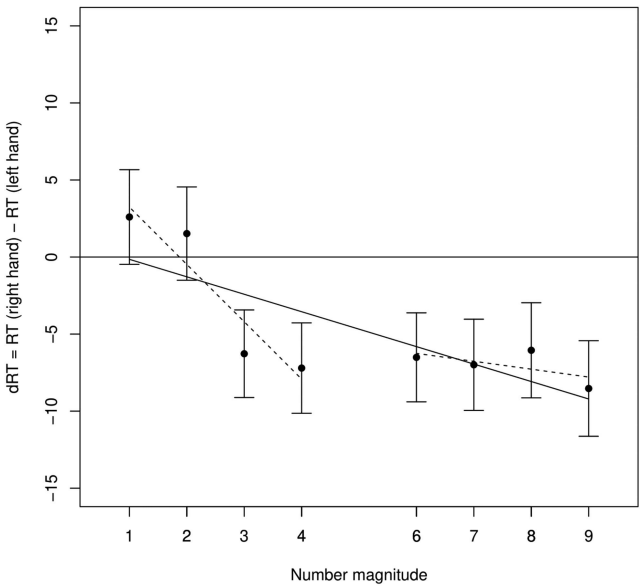


Note. Confidence intervals for the MARC effect (x-axis) for each participant separately (y-axis), with the empirically observed MARC slope marked as a dot. MARC = Linguistic Markedness of Response Codes; H0 = null hypothesis.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Figure 4

Overall dRTs per Number Magnitude in Color Intensity Judgment (Experiment 2)



Note. Dots represent dRTs averaged across all trials from all participants and reveal potential Spatial–Numerical Association of Response Codes (SNARC) and Linguistic Markedness of Response Codes (MARC) effects. Error bars represent ± 1 standard error for the respective number. Solid regression lines represent SNARC slopes for the whole stimulus set, while dashed regression lines represent SNARC slopes calculated separately for the lower range (number magnitudes 1, 2, 3, and 4) and for the higher range (number magnitudes 6, 7, 8, and 9). A present MARC effect would be reflected by dRTs being on average larger for odd than for even numbers. dRT = differences in response times between both hands; RT = response time.

Differences and Correlations of the SNARC and MARC Effects Between Blocks

No evidence was found for SNARC and MARC effects to differ between the first and second block in Experiment 1. While no evidence was found for the MARC effect to differ between blocks either, the SNARC effect was stronger in the first than in the second block in Experiment 2. Interestingly, no correlations of the SNARC and MARC effects between blocks were found. More detailed results are reported in Appendix I.

Range Dependency of the SNARC Effect

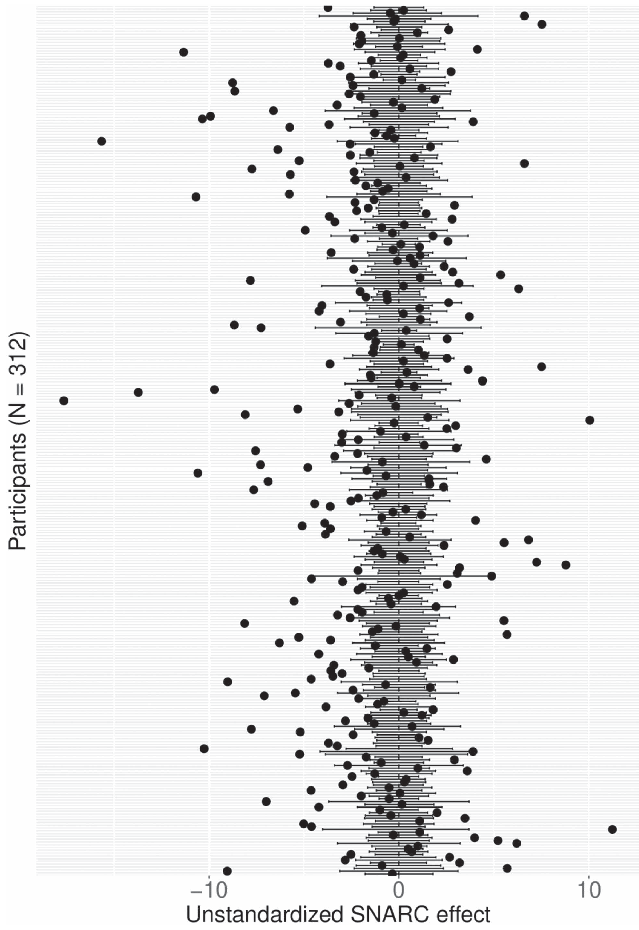
Last, when looking at Figures 1 and 4 displaying dRTs per number magnitude in Experiments 1 and 2, the SNARC effect seems to be driven by the small numbers in the stimulus set, while the large numbers do not seem to contribute. We therefore calculated the SNARC effect's range dependency by calculating SNARC slopes for the lower range (i.e., Numbers 1–4) and for the higher range (i.e., Numbers 6–9) separately and found them to differ, both in the full samples and in the subsamples showing an overall reliable SNARC effect. Details can be found in Appendix J.

Discussion

The present study aimed at estimating the extent to which the semantic number processing is automatic. Specifically, we investigated the degree of automaticity of processing magnitude and parity as well as the mental mapping of these two numerical features onto space. For this, so-called primitives (i.e., Arabic single-digit numbers, as defined by Tzelgov et al., 2015) were used as stimuli. Many previous studies have demonstrated the SNARC effect in tasks requiring semantic number processing, such as parity judgment, even when neither processing of number magnitude nor its mapping onto space are demanded. The SNARC effect has also been found in some tasks requiring only processing of nonsemantic number spatial features, such as orientation, suggesting automatic activation of number semantic and of other spatial codes when some sort of spatial processing is required (see neural overlap hypothesis by Fias et al., 2001). However, evidence from color judgment was mixed, and

Figure 5

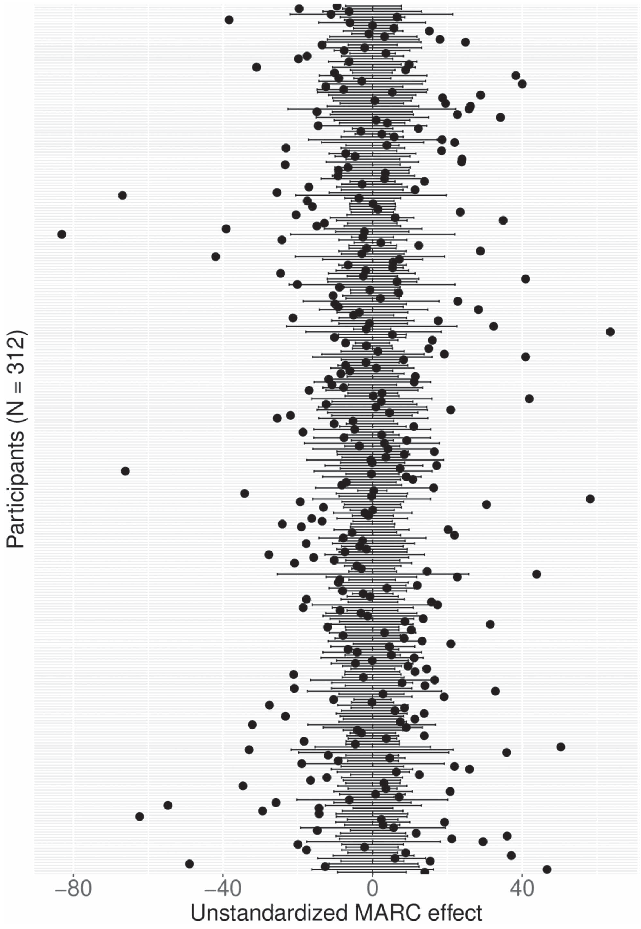
90% H0 Bootstrap Confidence Intervals for the SNARC Slopes in Color Intensity Judgment (Experiment 2)



Note. Confidence intervals for the SNARC effect (*x*-axis) for each participant separately (*y*-axis), with the empirically observed SNARC slope marked as a dot. Overall differences in response times between both hands per number magnitude for participants revealing a reliable regular SNARC effect are illustrated in Figure E2. SNARC = Spatial–Numerical Association of Response Codes; H0 = null hypothesis.

Figure 6

90% H0 Bootstrap Confidence Intervals for the MARC Slopes in Color Intensity Judgment (Experiment 2)



Note. Confidence intervals for the MARC effect (x-axis) for each participant separately (y-axis), with the empirically observed MARC slope marked as a dot. MARC = Linguistic Markedness of Response Codes; H0 = null hypothesis.

existing studies revealed limitations such as small sample sizes, few repetitions per experimental cell (note that at least 20 participants and 20 repetitions are required to detect the SNARC effect according to simulations run by Cipora & Wood, 2017), and insufficient statistical power to detect small effects. Additionally, some of the previous studies introduced confounds due to the colors used and lacked counterbalancing of experimental conditions. Therefore, we methodologically improved two often-quoted experiments to minimize design-related confounds and moved them to an online setting with large samples guaranteeing high statistical power for small effects. Namely, we replicated nominal color judgment and color intensity judgment. To make the present study as transparent as possible, the hypotheses and method were preregistered, and all data and analyses are publicly available. With these two highly transparent online experiments, we could shed light on the replicability and validity of previously reported results and previous claims about the (non)-automaticity of human numerical cognition, which relied partly on underpowered studies with certain confounds. The present study thus forms part of the growing body of replication studies in psychological and cognitive research, demonstrating how research methodology can be improved to obtain trustworthy results.

Automatic Number Magnitude Processing Reflected by the SNARC Effect

Crucially, a group-level SNARC effect was demonstrated both in nominal color judgment and in color intensity judgment, which points toward high automaticity of both the processing and the spatial mapping of number magnitude. This is in line with our hypotheses, but not with the results of the two original studies by Fias et al. (2001) and Lammertyn et al. (2002), nor with later studies by Schroeder et al. (2017a, 2017b), Cleland and Bull (2019), and Didino et al. (2019), who failed to detect the SNARC effect in these tasks. The SNARC effects in the present study were relatively small, with slopes between −2 and −1 ms. Previous null results can be explained by lacking statistical power for detecting such small SNARC effect sizes. One possible reason for the SNARC effect to be rather small is that RTs are overall faster in color judgment than in number-magnitude classification and parity judgment (Didino et al., 2019; Keus & Schwarz, 2005), and the SNARC effect is usually less pronounced in shorter RTs. However, a more convincing reason might be that color judgments require neither the processing of semantic numerical features nor of spatial features. We argue that the strength of the SNARC increases with increasing focus on number semantics and with increasing neural overlap between the response-relevant dimension and number magnitude. According to this claim and as summarized in Figure 7, color judgment elicits a weaker SNARC effect than orientation judgment or the judgment of semantic numerical features, and the SNARC effect should be strongest in tasks where both semantic numerical features and spatial features are response-relevant (such as in U. Fischer et al., 2018).

Finding a SNARC effect in both color judgment tasks has several important implications for existing theories of numerical cognition, even if the observed effect is small. First, the processing of number magnitude and its mapping onto space in human minds seem to be even more automatic than previously assumed. Until the present, consistent evidence for the SNARC effect in processing of nonsemantic number features came from orientation judgment of numbers or of superimposed objects (e.g., Fias et al., 2001; Lammertyn et al., 2002), while evidence from color judgment was rather inconsistent. The SNARC effect is weaker when only nonsemantic features of numbers need to be processed than when the semantic property of number magnitude is directly activated (Didino et al., 2019), but still evoked, as demonstrated in the present study. These findings can be explained by the instance theory (Logan, 1988): As single-digit numbers are instances stored in long-term memory and repeatedly retrieved by educated adults from there in everyday life, processing single-digit numbers is highly automatized in educated adults. According to Tzelgov et al. (2015), single-digit numbers can therefore also be called primitives. Moreover, the findings are in line with the mental number line account of the SNARC effect (Dehaene et al., 1993), which is assumed to be a spatial mental representation of numerical magnitude in long-term memory. According to this account, numerical codes are strongly intertwined with spatial magnitude codes in long-term

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Figure 7

Strength of the SNARC Effect Depending on the Response-Relevant Feature (i.e., Spatial and/or Semantic Numerical Features to Be Processed) Compared Between Several Tasks

Task	Response-relevant feature	SNARC effect
Color judgment (e.g., nominal: blue vs. yellow; intensity: dark cyan vs. light cyan)	Neither numerical features nor spatial features	Increasing SNARC effect with increasing focus on semantic numerical features or spatial features
Orientation judgment (e.g., number orientation: upright vs. tilted; triangles superimposed numbers: upwards vs. downwards; lines superimposed on numbers: horizontal vs. vertical)	Spatial feature but no numerical features	
Parity judgment (i.e., odd vs. even)	Numerical feature (other than number magnitude) but no spatial features	
Magnitude classification (e.g., smaller than 5 vs. larger than 5, when 5 is the fixed standard)	Numerical feature (i.e., number magnitude) but no spatial features	
Magnitude classification (e.g., smaller than 5 vs. larger than 5, when 5 is the fixed standard) with the fixed standard being presented on a number line (e.g., from 1 on the left to 9 on the right) along with each stimulus to be classified (as in Fischer et al., 2018)	Numerical feature (i.e., number magnitude) and spatial feature (i.e., location of the fixed standard on a number line)	

Note. SNARC = Spatial–Numerical Association of Response Codes.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

memory, leading to them being activated simultaneously and thus producing the SNARC effect.

Second, interesting conclusions can be drawn from the findings of the present study regarding the neural overlap hypothesis by Fias et al., claiming that the SNARC effect arises only in case of overlapping neural structures for processing relevant and irrelevant information. If different neural structures are involved in number and color processing (as stated by Fias et al.), our findings speak against the neural overlap hypothesis: Number magnitude should not be processed in color judgments, if separate neural pathways are utilized, and therefore, the SNARC effect could not be triggered in such tasks.

Interestingly, the SNARC effect is driven by the smaller numbers in both nominal color judgment and color intensity judgment. In other words, the SNARC slope between 1 and 4 is steeper than the SNARC slope between 6 and 9. This in line with findings from other tasks requiring only processing of nonsemantic number features. For instance, in orientation judgment of squares surrounding numbers, Yu et al. (2018) found a larger SNARC effect in a stimulus set from 1 to 6 than from 4 to 9. This pattern is not only typical in judgments of physical features of numerical stimuli or superimposed objects but also in judgments about semantic number features. In their seminal SNARC article, Dehaene et al. (1993) found a descriptively stronger SNARC effect between 0 and 5 than between 4 and 9 in the classical parity judgment task, although the difference remained nonsignificant (see also Roth et al., 2025). The range dependency suggests that processing of number magnitude and its mapping onto space is more automatic for small than for large numbers. Importantly, we cannot determine whether the SNARC effect was only driven by Numbers 1 and 2 or whether Number 3 was also part of the typical negative regression slope, because its dRT was similar to dRTs for larger numbers from 4 to 9. Another possibility is that the SNARC effect exists for the full stimulus set but that the increase in right- over left-hand advantage per increase in number magnitude changes in a logarithmic way (i.e., the change is largest from 1 to 2 and smallest from 8 to 9). One potential explanation for the SNARC effect being driven by smaller numbers might be that only the smallest primitives (i.e., 1, 2, and 3) are automatically associated with the left, whereas larger single-digit numbers beyond the subitizing range are not associated with the right in an equally automatic fashion. Alternatively, instead of the numbers 1, 2, 3 per se, the first, second, and third number of an experiment or given number range might be automatically mapped to space (i.e., a kind of spatial–numerical primacy effect). In our view, the observed data pattern for automaticity are highly interesting and relevant, and multiple follow-up experiments are conceivable. For instance, the SNARC effect must be investigated in color judgment with different

stimulus sets starting with 1, 2, and 3 or with 4, 5, and 6. If the automatic SNARC effect turns out to be more pronounced for the first than for the latter stimulus sets, the reason could be that small numbers are more often used in daily life than large numbers, which is why their mental representation might be more automatic. The fact that this typical feature of the SNARC effect arose in the present study suggests that the SNARC effect is qualitatively similar in judgments of nonsemantic and semantic number features. Note that Bayesian evidence regarding a potential difference in the SNARC effect's size between the color judgment tasks remained inconclusive. Also, the pattern of a stronger SNARC in smaller than in larger numbers was similar in both nominal color judgment and color intensity judgment.

Interestingly, we found strong Bayesian evidence against the size effect (Fias et al., 2003) in color intensity judgment, and we even observed the reversed pattern in nominal color judgment, with responses getting faster with increasing number magnitude. However, this pattern is largely driven by particularly slow responses to Number 1 in both experiments and extremely fast responses to Number 7 in color intensity judgment (see Figures D1 and D2). Fast responses for Number 7 have also been found in a number classification task by Cohen (2009). Importantly, we know that number magnitude was processed to some extent in both experiments, as this together with its association with space is a prerequisite for the observed SNARC effect (Cipora, He, et al., 2020). The size effect might be absent despite the presence of the SNARC effect, because number magnitude does not necessarily impact response speed; instead, it may only activate the corresponding mental representation, when number magnitude is irrelevant for the task at hand.

Automatic Number Parity Processing Reflected by the Odd and MARC Effects

Apart from automaticity of number magnitude processing and the SNARC effect, both color judgment tasks enabled us to investigate automaticity of number parity processing and the MARC effect. Most crucially, we replicated the odd effect (Hines, 1990) in nominal color judgment, indicating faster responses to even than to odd numbers (see Figure D1), both when analyzing it separately and when taking into account other main effects or interactions by retaining them in the LMM. In contrast, we found strong Bayesian evidence against the odd effect in color intensity judgment when analyzing it separately but could detect it when taking other effects into account. These results provide evidence for automatic processing of number parity in color judgment.

Apart from the odd effect, we found a significant group-level MARC effect in nominal color judgment, although the estimated increase of less than 3 ms in right- over left-hand advantage for even compared to odd numbers is rather small and Bayesian evidence remained inconclusive or was, at best, moderate. This effect disappeared in the LMM when simultaneously considering other effects. Also, it is important to note that the MARC effect might be partly or fully driven by the SNARC effect's range dependency. That is, the SNARC effect is only present between Numbers 1 and 3, where the average magnitude is lower for odd than for even numbers, but not between 4 and 9, where the average magnitude is lower for even than for odd numbers. This causes dRTs to be overall higher for odd than for even numbers, which is in turn reflected by more negative MARC estimates and thus contributes to a significant MARC effect. Note that number parity was included into the repeated-measures regression models along with number magnitude saying that these two predictors are orthogonal to each other, but this is not true in case of range dependency of the SNARC effect. In color intensity judgment, the group-level MARC effect remained nonsignificant, with strong Bayesian evidence against it; it remained nonsignificant when taking other effects into account. The difference between the two tasks was nonsignificant, with an inconclusive Bayes Factor. Importantly, previous research has not investigated the MARC effect as thoroughly as the SNARC effect and, to our knowledge, has not tested its presence in color judgment tasks yet. Nuerk et al. (2004) found the MARC effect to be more reliable and stronger for number words than for Arabic digits. Roettger and Domahs (2015, Experiment 1) replicated the MARC effect for number words, but only in judgments of semantic number features (e.g., parity judgment, magnitude classification) and not nonsemantic number features (e.g., font classification, lexical decision). The present findings align with the literature, demonstrating that the MARC effect emerges when semantic number processing is required but is very small or absent when only processing nonsemantic number features is necessary.

In summary, the processing of number parity (i.e., odd effect) seems to arise even when number parity is not relevant to the task, while its spatial mapping (i.e., MARC effect) is not as automatic as the spatial mapping of number magnitude (i.e., SNARC effect).

Individual Prevalence of the SNARC and MARC Effects in Color Judgments

Apart from investigating the SNARC and MARC effects at the group level, we also looked at their prevalence at the individual level by means of bootstrapping. In color judgment, 48% (Experiment 1) and 41% (Experiment 2) of the participants revealed a reliable (i.e., stronger than what would be likely considering intra-individual variations in RTs) SNARC effect. These proportions are similar and even slightly higher as compared to parity judgment (40% in Cipora, van Dijck, et al., 2019). Interestingly, 17% (Experiment 1) and 19% (Experiment 2) of the sample showed a reliable reversed SNARC effect, which clearly is a larger proportion than in parity judgment (4% in Cipora, van Dijck, et al., 2019). Regarding the prevalence of the MARC effect, 36% (Experiment 1) and 28% (Experiment 2) of the sample revealed the effect reliably, which are similar proportions as compared to parity judgment (35% in Cipora, van Dijck, et al., 2020). Crucially, 24% (Experiment 1) and 29% (Experiment 2) of the sample showed a reliable reversed MARC effect, which is again a larger proportion than in parity judgment (18% in Cipora, van Dijck, et al., 2020). The high prevalence of reliable reversed SNARC and MARC effects in both color judgment tasks is an unexpected observation, although the SNARC effect has been shown to be an indominant phenomenon (Cipora, van Dijck, et al., 2019; Roth et al., 2024; i.e., it is not present or even reversed in some individuals; Rouder & Haaf, 2018). These large proportions of participants showing a reversed SNARC effect are a possible explanation for the small SNARC and MARC effects at the group level. Namely, the regular effects might be equal in color judgment when compared to other tasks in terms of size and prevalence, but the higher prevalence of the reversed effects reduces the mean effect size at the group level. That is, although the present study revealed relatively small

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

SNARC effects for color judgment at the group level (i.e., slopes between −2 and −1 ms) as compared to SNARC effects in parity judgment (i.e., slopes typically around −5 ms), we therefore argue that the observed effects are equally meaningful in both kinds of tasks. Note that, in the case of parity judgment, the proportion of participants revealing the reliable reversed SNARC effect is close to the false positive rate allowed by a 90% confidence level, while this proportion is considerably higher here. One may even argue that in the case of color judgment, the automaticity of number magnitude processing and its mapping onto space is not lower than in the case of semantic number features. This is because the overall proportion of participants revealing nonrandom spatial mapping in color judgment is even *greater* than in a typical parity judgment setup. We observe greater variation in the direction of the spatial mapping. Kramer et al. (2018) argue that the reversed SNARC effect should be treated as evidence for automatic processing of number magnitude as well, just like the regular SNARC. Therefore, the conclusion that automaticity is lower, which can be derived from less pronounced group-level slopes, may not be correct, as it may simply be more inconsistent between participants.

As outlined in the introduction (Casasanto & Pitt, 2019; Pitt & Casasanto, 2022; Prpic et al., 2021; Walsh, 2003), not only numbers can be interpreted as a magnitude or ordinality system, but also luminance, which is the task-relevant feature in color intensity judgment. The luminance dimension, which was present in color intensity judgment but not in nominal color judgment, could have either increased the activation of the number magnitude dimension, thereby increasing the SNARC effect, or suppressed this activation, thereby reducing the SNARC effect. Neither of the two hypotheses was supported by our large amount of data, and no evidence for or against a difference in SNARC effects between nominal color judgment and color intensity judgment was found.

Response Patterns Depending on Sample Characteristics

The present experiments were designed as conceptual online replications of the studies by Fias et al. (2001) and Lammertyn et al. (2002). Responses in color intensity judgment were on average faster in our study (460 ms after applying exclusion criteria) as compared to the study by Fias et al. (between 481 and 513 ms in their Experiment 3) and Lammertyn et al. (2002; between 555 and 590 ms). This result is in line with our expectation, because Prolific users are probably more trained in speeded RT tasks. However, comparing absolute millisecond values between online and offline experiments is problematic, as there is a myriad of factors that can influence the absolute measurement (Garaizar & Reips, 2019). Surprisingly, for nominal color judgment, responses were slower in the present study (443 ms after applying exclusion criteria) than in the original study by Fias et al. (between 375 and 388 ms in their Experiment 2). A reason for this pattern might be that it was easier to judge the task-relevant feature of the stimuli in Fias et al.'s study, because stimuli did not only differ in their color hue but also in their luminance. Nevertheless, responses were descriptively faster in nominal color judgment than in color intensity judgment, which is consistent with the literature.

In the exploratory analyses (Appendices F, G, and H), we investigated potential relationships between demographic variables or other participant characteristics and the SNARC and MARC effects. First, we asked participants whether they were affected by synesthesia, because the color of the presented number might trigger another number that is associated with this color, or the presented number might trigger another color that is associated with this number. This can result in congruency effects, as described by Cohen Kadosh, Tzelgov, and Henik (2008). In the two synesthesia subsamples of both experiments, SNARC effects (inconclusive Bayes factors) and MARC effects (moderate Bayesian evidence against them) remained nonsignificant, but we found moderate Bayesian evidence against a difference between those participants who are affected by synesthesia and those who are not. As can be seen in Appendix H, the SNARC effects would have been detectable in the majority (Experiment 1) or almost half (Experiment 2) of the bootstrapped nonsynesthesia subsamples of the same size as the synesthesia subsample, but the MARC effects only in very few bootstrapped nonsynesthesia subsamples (Experiments 1 and 2). In any case, compared to several previous studies, a strength of our study was the counterbalanced design: Each digit was equally often presented in each color and with each response-to-key assignment. Hence, even if some participants were affected by digit-color synesthesia and associate yellow or cyan with any of the digits between 1 and 9, there should not be any confound leading to a differential pattern in these participants.

Next, we investigated the SNARC and MARC effects separately for left-handers (28 in each experiment) and compared them with right-handers (301 in Experiment 1 and 273 in Experiment 2). Although none of the effects were significant in left-handers, no difference was found as compared to the effects in right-handers (see Appendix H). When simultaneously taking into consideration other effects (e.g., the odd effect) in an LMM, however, the SNARC effect seemed to be strongest in right-handed, intermediate in ambidextrous, and weakest in left-handed participants in both experiments. In contrast, no difference in MARC effects was found in the LMMs for both experiments. In previous studies, no influence of handedness on the SNARC effect has been found (Cipora, Soltanlou, et al., 2019; Dehaene et al., 1993; Huber et al., 2015; Müller & Schwarz, 2007), whereas the MARC effect seems to exist only in right-handers but is absent or even reversed in left-handers (Cipora, Soltanlou, et al., 2019; Huber et al., 2015).

Moreover, exploratory analyses revealed moderate Bayesian evidence against correlations with math skills (in line with Cipora, He, et al., 2020) and with finger-counting habits and their stability (in line with Hohol et al., 2022). Also, we found strong Bayesian evidence against differences in the SNARC or MARC effects between native speakers of six different left-to-right written languages. Note that results might be different for native speakers of other languages (Bulut et al., 2024). Further, the data revealed moderate Bayesian evidence against a correlation of the unstandardized SNARC and MARC effect sizes between the two halves of both experiments. Also, we found moderate Bayesian evidence against differences in the MARC effect between the two halves of both experiments as well as in the SNARC effect in nominal color judgment. In contrast, the unstandardized SNARC effect was significantly stronger in the first half than in the second half of color intensity judgment, although the Bayes Factor remained inconclusive. A slight decrease in the SNARC effect across the experiment could be explained by responses getting faster across the experiment and the SNARC

effect typically being stronger in slower than in faster responses (Didino et al., 2019).

Response Patterns Depending on Luminance, Magnitude, and Parity

The exploratory LMM revealed interesting response patterns depending on luminance color intensity judgment (see Appendix F). Note that we attempted to control for background luminance by using the exact average of the light and dark number stimuli in terms of CIELAB values, because Fumarola et al. (2014) argued that the background luminance might influence the spatial mapping of the stimuli's luminance. However, luminance can never be entirely controlled for, especially in online studies, where each participant runs the experiment on another device and in another environment. Additionally, screen reflections (Ramasoot & Fotios, 2012) and surrounding luminance (Sheedy et al., 2005) can play a role. First, participants responded faster to light than to dark cyan stimuli (as in Wang et al., 2022, with red stimuli; but see Wang et al., 2022 and Ren et al., 2011, with achromatic stimuli). Importantly, this result was not confirmed when interactions were not taken into consideration, with strong Bayesian evidence against a difference between light and dark cyan stimuli (see Appendix G). In nominal color judgment, we found moderate Bayesian evidence against a difference between yellow and cyan stimuli. Multiple luminance levels could be used for each color hue in future experiments involving nominal color judgment, so that luminance cannot be used as a source of information. Note, however, that we are not aware of any study showing that display parameters, such as overall luminances, contrasts, or similar visual features of number stimuli or background affect the SNARC effect. Second, our data collected with chromatic colors revealed an association of light/dark with the left/right, respectively, which corresponds to the association found by Ren et al. (2011) in achromatic stimuli, whereas Fumarola et al. (2014), Gusmão et al. (2022), Löffler et al. (2022), and Wang et al. (2022) reported the reversed association in chromatic stimuli. The present finding is highly surprising, and we cannot find any explanation in the literature, nor can we come up with an ad hoc explanation for the direction of this association. Nevertheless, the present results should be taken seriously, because this is the most highly powered study in the field.

Beyond the unexpected direction of the luminance–space interaction, the findings support the hypothesis that, on top of numbers, other magnitude or ordinality dimensions (e.g., color intensity) are also mapped onto space (i.e., spatial quantity association of response codes effect; Löffler et al., 2022), in line with a theory of magnitude (Walsh, 2003) or the ordinal position effect (i.e., working memory account; van Dijck et al., 2014). Further, an association of small/large numbers with high/low luminance was found, respectively. This finding is in line with Cohen Kadosh and Henik (2006) and Cohen Kadosh et al. (2007), but stands in contrast to the reversed association found by Cohen Kadosh, Cohen Kadosh, and Henik (2008) and Pinel et al. (2004). Note that the association of high/low luminance with small/large numbers is closely linked to the above-described association of high/low luminance with the left/right side, because the SNARC effect describes the association of small/large numbers with the left/right side. Note that this was not the central question of the present study and needs further resolution, both empirically and theoretically.

Conclusion

Humans can process numbers either intentionally when required in a given situation or automatically, even when not necessary for the task at hand. The SNARC effect reflects that number magnitude is processed and mentally mapped onto space automatically, even when number processing is not necessary, at least for small numbers like 1, 2, and possibly also 3. We know that the SNARC effect arises in judgments on semantic features of numbers other than their magnitude, but evidence from judgments on nonsemantic features of numbers such as their color is mixed. In the present study, which importantly improves on many past studies in realizing adequate power via online experimentation (Cipora, Soltanlou, et al., 2019; Reips, 2007, 2021), we demonstrated the presence of a small SNARC effect in nominal color judgment as well as in color intensity judgment with high statistical power in two large online experiments while controlling for confounding variables as much as possible. At the same time, little to no evidence for a small MARC effect was found in nominal color judgment and in color intensity judgment, although we found evidence for the odd effect in both tasks.

In summary, the processing of number magnitude and of number parity seem to be highly automatic, while the spatial mapping of number magnitude seems to be more automatic than the spatial mapping of number parity. Even if the observed effects in the present study are rather small, they are highly important for understanding the nature of number processing, because they demonstrate highly automatized processing of number magnitude. In a broader context, this study demonstrates how research can be made transparent and how methodologically and theoretically sound replication studies can be conducted. Although the observed effects are small in the present study, they show that conclusions drawn about the lack of automaticity in semantic number processing, derived from frequentist null results in underpowered studies, are not valid. The degree of automaticity in human cognition might therefore be higher than previously assumed.

References

Avitan, A., Wasserman, S., & Henik, A. (2024). Endogenous attention modulates automaticity of number processing. *Psychonomic Bulletin & Review*, *31*(4), 1579–1587. https://doi.org/10.3758/s13423-023-02438-4

Bulut, M., Roth, L., Bahreini, N., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2024). One direction? Cultural aspects of the mental number line beyond reading direction. *Psychological Research*, *89*, Article 37. https://doi.org/10.1007/s00426-024-02038-4

Casasanto, D., & Pitt, B. (2019). The faulty magnitude detector: Why SNARC-like tasks cannot support a generalized magnitude system. *Cognitive Science*, *43*(10), Article e12794. https://doi.org/10.1111/cogs.12794

Cipora, K., He, Y., & Nuerk, H.-C. (2020). The spatial–numerical association of response codes effect and math skills: Why related? *Annals of the New York Academy of Sciences*, *1477*(1), 5–19. https://doi.org/10.1111/nyas.14355

Cipora, K., & Nuerk, H.-C. (2023). What the attentional-SNARC and its (null) replications can and cannot tell us. *Meta-Psychology*, *7*. https://doi.org/10.15626/MP.2020.2598

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Cipora, K., Patro, K., & Nuerk, H.-C. (2018). Situated influences on spatial–numerical associations. In T. L. Hubbard (Ed.), *Spatial biases in perception and cognition* (pp. 41–59). Cambridge University Press. https://doi.org/10.1017/9781316651247.004

Cipora, K., Soltanlou, M., Reips, U.-D., & Nuerk, H.-C. (2019). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behavior Research Methods*, *51*(4), 1676–1692. https://doi.org/10.3758/s13428-019-01213-5

Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Goebel, S. M., Willmes, K., Pesenti, M., Schiltz, C., & Nuerk, H.-C. (2019). *A minority pulls the sample mean: On the individual prevalence of robust group-level cognitive phenomena—The instance of the SNARC effect* [Preprint]. Open Science Framework. https://doi.org/10.31234/osf.io/bwyr3

Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Schiltz, C., Pesenti, M., Goebel, S. M., Willmes, K., & Nuerk, H.-C. (2020). *Parity and space: On the prevalence and individual differences in the MARC effect* [Poster presentation]. XXXVIIIth European Workshop on Cognitive Neuropsychology, Bressanone, Italy. https://osf.io/4qhpw

Cipora, K., & Wood, G. (2017). Finding the SNARC instead of hunting it: A 20*20 Monte Carlo investigation. *Frontiers in Psychology*, *8*, Article 1194. https://doi.org/10.3389/fpsyg.2017.01194

Claeys, K. G., Dupont, P., Cornette, L., Sunaert, S., Van Hecke, P., De Schutter, E., & Orban, G. A. (2004). Color discrimination involves ventral and dorsal stream visual areas. *Cerebral Cortex*, *14*(7), 803–822. https://doi.org/10.1093/cercor/bhh040

Cleland, A. A., & Bull, R. (2019). Automaticity of access to numerical magnitude and its spatial associations: The role of task and number representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(2), 333–348. https://doi.org/10.1037/xlm0000590

Cleland, A. A., Corsico, K., White, K., & Bull, R. (2020). Non-symbolic numerosities do not automatically activate spatial–numerical associations: Evidence from the SNARC effect. *Quarterly Journal of Experimental Psychology*, *73*(2), 295–308. https://doi.org/10.1177/1747021819875021

Cohen, D. J. (2009). Integers do not automatically activate their quantity representation. *Psychonomic Bulletin & Review*, *16*(2), 332–336. https://doi.org/10.3758/PBR.16.2.332

Cohen Kadosh, R., Cohen Kadosh, K., & Henik, A. (2008). When brightness counts: The neuronal correlate of numerical-luminance interference. *Cerebral Cortex*, *18*(2), 337–343. https://doi.org/10.1093/cercor/bhm058

Cohen Kadosh, R., & Henik, A. (2006). A common representation for semantic and physical properties: A cognitive-anatomical approach. *Experimental Psychology*, *53*(2), 87–94. https://doi.org/10.1027/1618-3169.53.2.87

Cohen Kadosh, R., Henik, A., Rubinsten, O., Mohr, H., Dori, H., van de Ven, V., Zorzi, M., Hendler, T., Goebel, R., & Linden, D. E. J. (2005). Are numbers special? The comparison systems of the human brain investigated by fMRI. *Neuropsychologia*, *43*(9), 1238–1248. https://doi.org/10.1016/j.neuropsychologia.2004.12.017

Cohen Kadosh, R., Henik, A., & Walsh, V. (2007). Small is bright and big is dark in synaesthesia. *Current Biology*, *17*(19), R834–R835. https://doi.org/10.1016/j.cub.2007.07.048

Cohen Kadosh, R., Lammertyn, J., & Izard, V. (2008). Are numbers special? An overview of chronometric, neuroimaging, developmental and comparative studies of magnitude representation. *Progress in Neurobiology*, *84*(2), 132–147. https://doi.org/10.1016/j.pneurobio.2007.11.001

Cohen Kadosh, R., Tzelgov, J., & Henik, A. (2008). A synesthetic walk on the mental number line: The size effect. *Cognition*, *106*(1), 548–557. https://doi.org/10.1016/j.cognition.2006.12.007

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1–2), 1–42. https://doi.org/10.1016/0010-0277(92)90049-N

Dehaene, S., & Akhavein, R. (1995). Attention, automaticity, and levels of representation in number processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 314–326. https://doi.org/10.1037/0278-7393.21.2.314

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371–396. https://doi.org/10.1037/0096-3445.122.3.371

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F., & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, *395*(6702), 597–600. https://doi.org/10.1038/26967

Didino, D., Breil, C., & Knops, A. (2019). The influence of semantic processing and response latency on the SNARC effect. *Acta Psychologica*, *196*, 75–86. https://doi.org/10.1016/j.actpsy.2019.04.008

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice*, *8*(1), 9–26. https://doi.org/10.1037/cns0000258

Elias, L. J., Saucier, D. M., Hardie, C., & Sarty, G. E. (2003). Dissociating semantic and perceptual components of synaesthesia: Behavioural and functional neuroanatomical investigations. *Cognitive Brain Research*, *16*(2), 232–237. https://doi.org/10.1016/S0926-6410(02)00278-1

Fias, W., Brysbaert, M., Geypens, F., & D'Ydewalle, G. (1996). The importance of magnitude information in numerical processing: Evidence from the SNARC effect. *Mathematical Cognition*, *2*(1), 95–110. https://doi.org/10.1080/135467996387552

Fias, W., Lammertyn, J., Reynvoet, B., Dupont, P., & Orban, G. A. (2003). Parietal representation of symbolic and nonsymbolic magnitude. *Journal of Cognitive Neuroscience*, *15*(1), 47–56. https://doi.org/10.1162/089892903321107819

Fias, W., Lauwereyns, J., & Lammertyn, J. (2001). Irrelevant digits affect feature-based attention depending on the overlap of neural circuits. *Cognitive Brain Research*, *12*(3), 415–423. https://doi.org/10.1016/S0926-6410(01)00078-7

Fischer, M. H., & Shaki, S. (2014). Spatial associations in numerical cognition—From single digits to arithmetic. *Quarterly Journal of Experimental Psychology*, *67*(8), 1461–1483. https://doi.org/10.1080/17470218.2014.927515

Fischer, U., Huber, S., Nuerk, H.-C., Cress, U., & Moeller, K. (2018). Spatial presentations, but not response formats influence Spatial–Numerical Associations in adults. *Frontiers in Psychology*, *9*, Article 2608. https://doi.org/10.3389/fpsyg.2018.02608

Fumarola, A., Prpic, V., Da Pos, O., Murgia, M., Umiltà, C., & Agostini, T. (2014). Automatic spatial association for luminance. *Attention, Perception, & Psychophysics*, *76*(3), 759–765. https://doi.org/10.3758/s13414-013-0614-y

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Garaizar, P., & Reips, U.-D. (2019). Best practices: Two web-browser-based methods for stimulus presentation in behavioral experiments with high-resolution timing requirements. *Behavior Research Methods*, *51*(3), 1441–1453. https://doi.org/10.3758/s13428-018-1126-4

Gusmão, B., Löffler, C. S., & Topolinski, S. (2022). The impact of horizontal and vertical luminance SNARC compatibility on affective judgments. *Cognition and Emotion*, *36*(8), 1522–1530. https://doi.org/10.1080/02699931.2022.2132221

Hines, T. M. (1990). An odd effect: Lengthened reaction times for judgments about odd digits. *Memory & Cognition*, *18*(1), 40–46. https://doi.org/10.3758/BF03202644

Hoffmann, D., Hornung, C., Martin, R., & Schiltz, C. (2013). Developing number-space associations: SNARC effects using a color discrimination task in 5-year-olds. *Journal of Experimental Child Psychology*, *116*(4), 775–791. https://doi.org/10.1016/j.jecp.2013.07.013

Hohol, M., Wołoszyn, K., & Cipora, K. (2022). No fingers, no SNARC? Neither the finger counting starting hand, nor its stability robustly affect the

SNARC effect. *Acta Psychologica*, *230*, Article 103765. https://doi.org/10.1016/j.actpsy.2022.103765

Huber, S., Klein, E., Graf, M., Nuerk, H.-C., Moeller, K., & Willmes, K. (2015). Embodied markedness of parity? Examining handedness effects on parity judgments. *Psychological Research*, *79*(6), 963–977. https://doi.org/10.1007/s00426-014-0626-9

Keus, I. M. (2006). *Shedding light on the SNARC: Illuminating the association between numerical and spatial information* [Doctoral dissertation, Radboud Universiteit Nijmegen]. https://repository.ubn.ru.nl/handle/2066/56964

Keus, I. M., Jenks, K. M., & Schwarz, W. (2005). Psychophysiological evidence that the SNARC effect has its functional locus in a response selection stage. *Cognitive Brain Research*, *24*(1), 48–56. https://doi.org/10.1016/j.cogbrainres.2004.12.005

Keus, I. M., & Schwarz, W. (2005). Searching for the functional locus of the SNARC effect: Evidence for a response-related origin. *Memory & Cognition*, *33*(4), 681–695. https://doi.org/10.3758/BF03195335

Krajcsi, A., Lengyel, G., & Laczkó, Á. (2018). Interference between number magnitude and parity. *Experimental Psychology*, *65*(2), 71–83. https://doi.org/10.1027/1618-3169/a000394

Kramer, P., Bressan, P., & Grassi, M. (2018). The SNARC effect is associated with worse mathematical intelligence and poorer time estimation. *Royal Society Open Science*, *5*(8), Article 172362. https://doi.org/10.1098/rsos.172362

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). *lmerTest package: Tests in linear mixed effects models* (Version 3.1.3, R package) [Computer software]. https://www.jstatsoft.org/article/view/v082i13

Lammertyn, J., Fias, W., & Lauwereyns, J. (2002). Semantic influences on feature-based attention due to overlap of neural circuits. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *38*(5), 878–882. https://doi.org/10.1016/S0010-9452(08)70061-3

Landy, D. H., Jones, E. L., & Hummel, J. E. (2008). *Why spatial-numeric associations aren't evidence for a mental number line*. https://philpapers.org/rec/LANWSA

Lengyel, G., & Krajcsi, A. (2024). *The SNARC effect emerges only with symbolic numbers* [Preprint]. Open Science Framework. https://doi.org/10.31234/osf.io/2c9kr

Löffler, C. S., Gerten, J., Mamporia, M., Müller, J., Neu, T., Rumpf, J., Schiller, M., Schneider, Y., Wozniak, M., & Topolinski, S. (2022). Bright on the right feels right: SQUARC compatibility is hedonically marked. *Cognition and Emotion*, *36*(4), 767–772. https://doi.org/10.1080/02699931.2022.2053660

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492

Lorch, R. F., Jr., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 149–157. https://doi.org/10.1037/0278-7393.16.1.149

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2022). *BayesFactor: Computation of Bayes factors for common designs* (Version 0.9.12-4.4, R package) [Computer software]. https://richarddmorey.github.io/BayesFactor/

Müller, D., & Schwarz, W. (2007). Is there an internal association of numbers to hands? The task set influences the nature of the SNARC effect. *Memory & Cognition*, *35*(5), 1151–1161. https://doi.org/10.3758/BF03193485

Naccache, L., & Dehaene, S. (2001a). The priming method: Imaging unconscious repetition priming reveals an abstract representation of number in the parietal lobes. *Cerebral Cortex*, *11*(10), 966–974. https://doi.org/10.1093/cercor/11.10.966

Naccache, L., & Dehaene, S. (2001b). Unconscious semantic priming extends to novel unseen stimuli. *Cognition*, *80*(3), 215–229. https://doi.org/10.1016/S0010-0277(00)00139-6

Nuerk, H.-C., Iversen, W., & Willmes, K. (2004). Notational modulation of the SNARC and the MARC (linguistic markedness of response codes) effect. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *57*(5), 835–863. https://doi.org/10.1080/02724980343000512

Patro, K., Nuerk, H.-C., Cress, U., & Haman, M. (2014). How number-space relationships are assessed before formal schooling: A taxonomy proposal. *Frontiers in Psychology*, *5*, Article 419. https://doi.org/10.3389/fpsyg.2014.00419

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron*, *41*(6), 983–993. https://doi.org/10.1016/S0896-6273(04)00107-2

Pitt, B., & Casasanto, D. (2022). The order of magnitude: Why SNARC-like tasks (still) cannot support a generalized magnitude system. *Cognitive Science*, *46*(2), Article e13108. https://doi.org/10.1111/cogs.13108

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, *132*(3), 416–442. https://doi.org/10.1037/0033-2909.132.3.416

Prpic, V., Mingolo, S., Agostini, T., & Murgia, M. (2021). Magnitude and order are both elevant in SNARC and SNARC-like effects: A commentary on Casasanto and Pitt (2019). *Cognitive Science*, *45*(7), Article e13006. https://doi.org/10.1111/cogs.13006

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.0.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ramasoot, T., & Fotios, S. (2012). Lighting and display screens: Models for predicting luminance limits and disturbance. *Lighting Research & Technology*, *44*(2), 197–223. https://doi.org/10.1177/1477153511413354

Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, *49*, 243–256. https://doi.org/10.1026/1618-3169.49.4.243

Reips, U.-D. (2007). The methodology of Internet-based experiments. In A. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 373–390). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199561803.013.0024

Reips, U.-D. (2009). Internet experiments: Methods, guidelines, metadata. In B. E. Rogowitz & T. N. Pappas (Eds.), *SPIE proceedings, human vision and electronic imaging XIV* (p. 724008). International Society for Optics and Photonics. https://doi.org/10.1117/12.823416

Reips, U.-D. (2021). Web-based research in psychology. *Zeitschrift für Psychologie*, *229*(4), 198–213. https://doi.org/10.1027/2151-2604/a000475

Reips, U.-D., & Neuhaus, C. (2002). WEXTOR: A web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, & Computers*, *34*(2), 234–240. https://doi.org/10.3758/bf03195449

Ren, P., Nicholls, M. E. R., Ma, Y. Y., & Chen, L. (2011). Size matters: Non-numerical magnitude affects the spatial coding of response. *PLOS ONE*, *6*(8), Article e23553. https://doi.org/10.1371/journal.pone.0023553

Restle, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology*, *83*(2, Pt. 1), 274–278. https://doi.org/10.1037/h0028573

Rich, A. N., Bradshaw, J. L., & Mattingley, J. B. (2005). A systematic, large-scale study of synaesthesia: Implications for the role of early experience in lexical-colour associations. *Cognition*, *98*(1), 53–84. https://doi.org/10.1016/j.cognition.2004.11.003

Richter, M., & Wühr, P. (2022). The nature of associations between physical stimulus size and left–right response codes. *Journal of Cognition*, *5*(1), Article 15. https://doi.org/10.5334/joc.206

Roettger, T. B., & Domahs, F. (2015). Grammatical number elicits SNARC and MARC effects as a function of task demands. *The Quarterly Journal*

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

of Experimental Psychology, *68*(6), 1231–1248. https://doi.org/10.1080/17470218.2014.979843

Roth, L., Caffier, J., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2021). *True colors SNARC: Semantic number processing is highly automatic*. Open Science Framework. https://osf.io/xpyf7/

Roth, L., Caffier, J., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2022a). *SNARC automaticity: Categorical color jugment (blue vs. yellow)* [Preregistration]. Open Science Framework. https://doi.org/10.17605/OSF.IO/F2GB8

Roth, L., Caffier, J., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2022b). *SNARC automaticity: Color intensity jugment (light cyan vs. dark cyan)* [Preregistration]. Open Science Framework. https://doi.org/10.17605/OSF.IO/VBA7N

Roth, L., Caffier, J., Reips, U.-D., Nuerk, H.-C., Overlander, A. T., & Cipora, K. (2025). One and only SNARC? Spatial–Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude. *Royal Society Open Science*, *12*(1), Article 241585. https://doi.org/10.1098/rsos.241585

Roth, L., Jordan, V., Schwarz, S., Willmes, K., Nuerk, H.-C., van Dijck, J.-P., & Cipora, K. (2024). Don't SNARC me now! Intraindividual variability of cognitive phenomena—Insights from the Ironman paradigm. *Cognition*, *248*, Article 105781. https://doi.org/10.1016/j.cognition.2024.105781

Rouder, J. N., & Haaf, J. N. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19–26. https://doi.org/10.1177/2515245917745058

Schaefer, J., Opgen-Rhein, R., & Strimmer, K. (2021). *GeneNet: Modeling and inferring gene networks* (Version 1.2.16, R package) [Computer software]. https://cran.r-project.org/web/packages/GeneNet/index.html

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017a). Prefrontal neuromodulation reverses spatial associations of non-numerical sequences, but not numbers. *Biological Psychology*, *128*, 39–49. https://doi.org/10.1016/j.biopsycho.2017.07.008

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017b). Space in numerical and ordinal information: A common construct? *Journal of Numerical Cognition*, *3*(2), 164–181. https://doi.org/10.5964/jnc.v3i2.40

Seshan, V. E., & Whiting, K. (2023). *clinfun: Clinical trial design and data analysis functions* (Version 1.1.3, R package) [Computer software]. https://cran.r-project.org/web/packages/clinfun/index.html

Sheedy, J. E., Smith, R., & Hayes, J. (2005). Visual effects of the luminance surrounding a computer display. *Ergonomics*, *48*(9), 1114–1128. https://doi.org/10.1080/00140130500208414

Simner, J., Mulvenna, C., Sagiv, N., Tsakanikos, E., Witherby, S. A., Fraser, C., Scott, K., & Ward, J. (2006). Synaesthesia: The prevalence of atypical cross-modal experiences. *Perception*, *35*(8), 1024–1033. https://doi.org/10.1068/p5469

Simon, J. R. (1990). The effects of an irrelevant directional cue on human information processing. In R. W. Proctor & T. G. Reeve (Eds.), *Advances in psychology: Stimulus–response compatibility* (Vol. 65, pp. 31–86). North-Holland. https://doi.org/10.1016/S0166-4115(08)61218-2

Toomarian, E. Y., & Hubbard, E. M. (2018). On the genesis of spatial–numerical associations: Evolutionary and cultural factors co-construct the mental number line. *Neuroscience and Biobehavioral Reviews*, *90*, 184–199. https://doi.org/10.1016/j.neubiorev.2018.04.010

Tzelgov, J., Ganor-Stern, D., Kallai, A. Y., & Pinhas, M. (2015). Primitives and non-primitives of numerical representations. In R. C. Kadosh & A. Dowker (Eds.), *Oxford library of psychology. The Oxford handbook of numerical cognition* (pp. 45–66). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.019

Tzelgov, J., Meyer, J., & Henik, A. (1992). Automatic and intentional processing of numerical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(1), 166–179. https://doi.org/10.1037/0278-7393.18.1.166

Tzelgov, J., Zohar-Shai, B., & Nuerk, H.-C. (2013). On defining quantifying and measuring the SNARC effect. *Frontiers in Psychology*, *4*, Article 302. https://doi.org/10.3389/fpsyg.2013.00302

van Dijck, J.-P., Abrahamse, E. L., Acar, F., Ketels, B., & Fias, W. (2014). A working memory account of the interaction between numbers and spatial attention. *The Quarterly Journal of Experimental Psychology*, *67*(8), 1500–1513. https://doi.org/10.1080/17470218.2014.903984

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. https://doi.org/10.1016/j.tics.2003.09.002

Wang, Q., Wu, Y., Zhang, F., Tao, W., & Wang, C. (2022). Processing directional information in stimuli inhibits the spatial association of luminance levels. *Attention, Perception, & Psychophysics*, *84*(8), 2596–2606. https://doi.org/10.3758/s13414-022-02572-6

Wickham, H. (2022). *plyr: Tools for splitting, applying and combining data* (Version 1.8.8, R package) [Computer software]. https://cran.r-project.org/web/packages/plyr/index.html

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2023). *ggplot2: Create elegant data visualisations using the grammar of graphics* (Version 3.4.3, R package) [Computer software]. https://cloud.r-project.org/web/packages/ggplot2/index.html

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* (Version 1.1.2, R package) [Computer software]. https://cran.r-project.org/web/packages/dplyr/index.html

Wickham, H., Vaughan, D., Girlich, M., & Ushey, K. (2023). *tidyr: Tidy messy data* (Version 1.3.0, R package) [Computer software]. https://cran.r-project.org/web/packages/tidyr/index.html

Wong, B. (2011). Color blindness. *Nature Methods*, *8*(6), Article 441. https://doi.org/10.1038/nmeth.1618

Wood, G., Willmes, K., Nuerk, H.-C., & Fischer, M. H. (2008). On the cognitive link between space and number: A meta-analysis of the SNARC effect. *Psychology Science*, *50*(4), 489–525.

Wühr, P., & Richter, M. (2022). Relative, not absolute, stimulus size is responsible for a correspondence effect between physical stimulus size and left/right responses. *Attention, Perception, & Psychophysics*, *84*(4), 1342–1358. https://doi.org/10.3758/s13414-022-02490-7

Yu, S., Li, B., Zhang, S., Yang, T., Jiang, T., Chen, C., & Dong, Q. (2018). Does the spatial–numerical association of response codes effect depend on digits' relative or absolute magnitude? Evidence from a perceptual orientation judgment task. *Journal of General Psychology*, *145*(4), 415–430. https://doi.org/10.1080/00221309.2018.1532391

(*Appendices follow*)

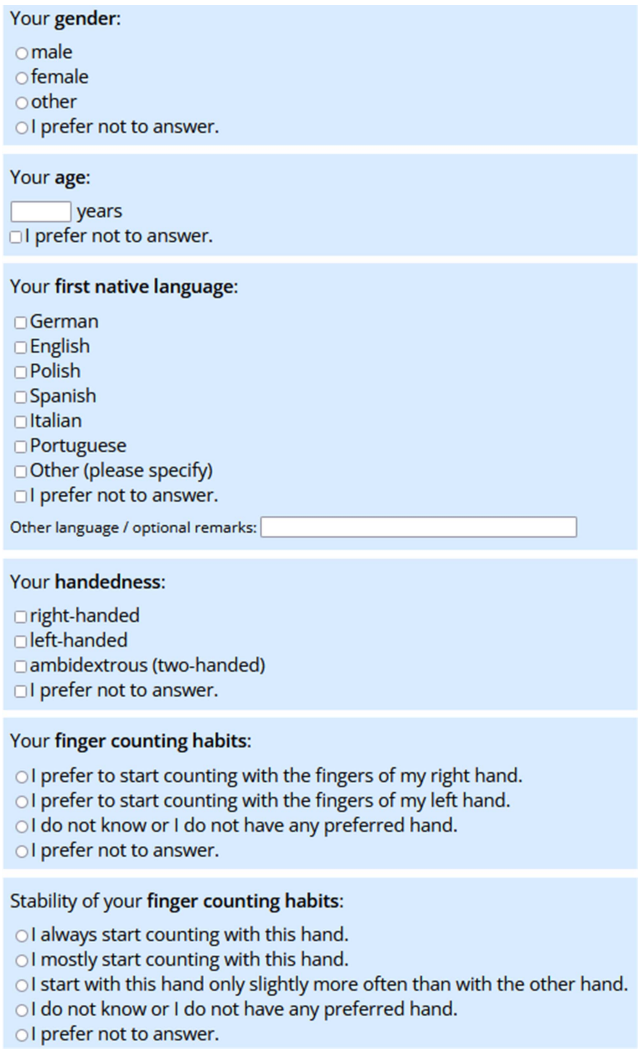
This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix A

Original Items for Descriptive Information About the Participants

Figure A1

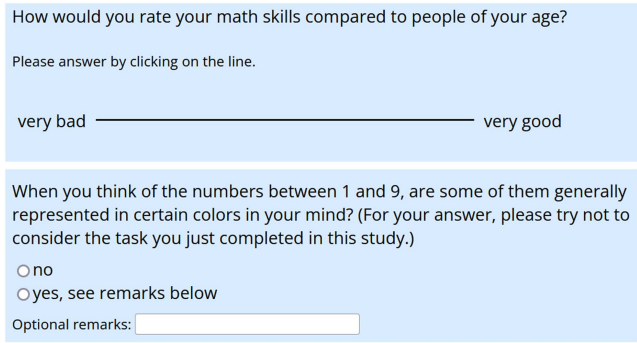
Six Original Items to Assess Demographic Information



Note. Screenshot of original demographic items that were presented before the color judgment task in both Experiments 1 and 2. See the online article for the color version of this figure.

Figure A2

Two Original Items to Assess Math Skills and Synesthesia



Note. Screenshot of original items that were presented after the color judgment task in both Experiments 1 and 2. See the online article for the color version of this figure.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix B

Color Vision Check and Color Judgment Tasks

Figure B1

Color Vision Check Prior to the Experimental Tasks

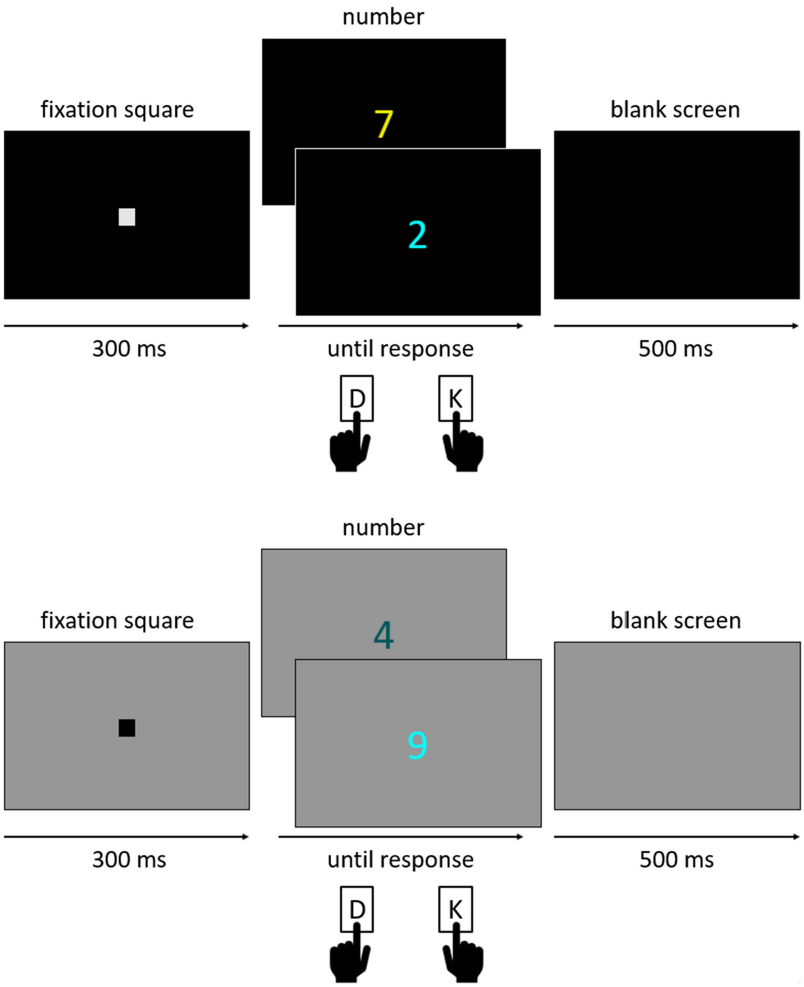
Note. Screenshots of original color vision checks prior to Experiment 1 (two upper screenshots) and prior to Experiment 2 (two lower screenshots), each consisting of two pages. See the online article for the color version of this figure.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Figure B2

Time Course of the Experimental Trials



Note. Two trials per experiment as examples: In Experiment 1 (upper part), the gray fixation square is followed by a yellow or cyan number on a black background, while in Experiment 2 (lower part), the black fixation square is followed by a dark or light cyan number on a gray background. Most participants used the D and K keys on their computer keyboard to respond to the presented numbers. See the online article for the color version of this figure.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix C

H0 Bootstrapping Confidence Intervals

Table C1

Proportion of Participants (%) Revealing a Reliable Regular SNARC Effect (Slope <0 ms), a Reliable Reversed SNARC Effect (Slope > 0 ms), or no Reliable Effect (Inconsistent) at Different Confidence Levels

Confidence level	SNARC effect	Experiment 1: Nominal color judgment	Experiment 2: Color intensity judgment
80%	Reliable regular	52.06	47.44
	Reliable reversed	20.59	21.15
	Inconsistent	27.35	31.41
95%	Reliable regular	42.94	36.86
	Reliable reversed	15.29	16.99
	Inconsistent	41.76	46.15
99%	Reliable regular	35.29	29.49
	Reliable reversed	11.76	13.14
	Inconsistent	52.94	57.37

Note. SNARC = Spatial–Numerical Association of Response Codes.

Table C2

Proportion of Participants (%) Revealing a Reliable Regular MARC Effect (Slope < 0 ms), a Reliable Reversed MARC Effect (Slope > 0 ms), or no Reliable Effect (Inconsistent) at Different Confidence Levels

Confidence level	MARC effect	Experiment 1: Nominal color judgment	Experiment 2: Color intensity judgment
80%	Reliable regular	40.00	32.05
	Reliable reversed	27.65	31.41
	Inconsistent	32.35	36.54
95%	Reliable regular	32.65	24.68
	Reliable reversed	19.12	26.60
	Inconsistent	48.24	48.72
99%	Reliable regular	27.65	20.19
	Reliable reversed	15.00	19.55
	Inconsistent	57.35	60.26

Note. MARC = Linguistic Markedness of Response Codes.

(*Appendices continue*)

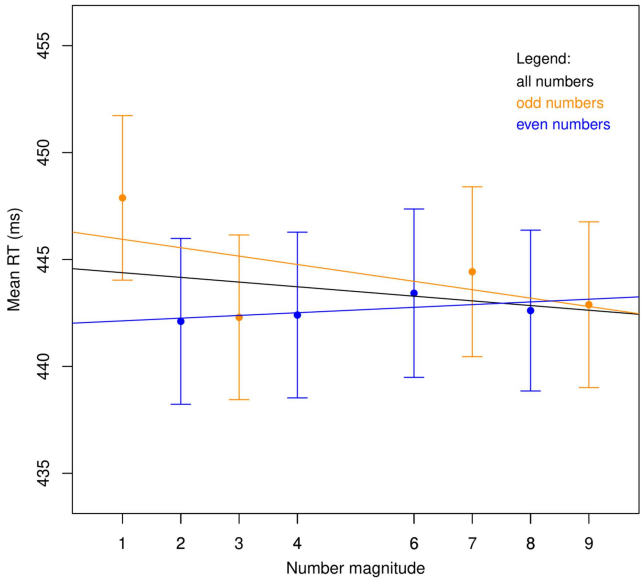
This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix D

Overall RTs for All Participants

Figure D1

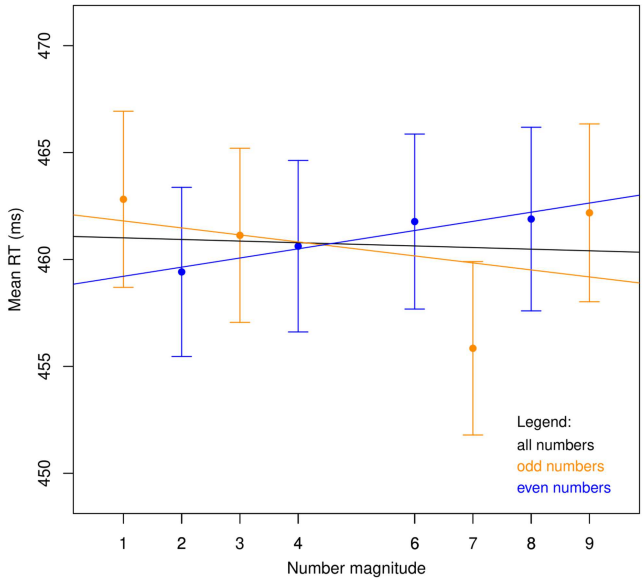
Overall RTs per Number Magnitude in Nominal Color Judgment (Experiment 1)



Note. Overall RTs per number magnitude averaged across all trials from all participants, with error bars representing ±1 standard error, with a regression line representing slope estimates for the odd numbers (1, 3, 7, and 9; color-coded in orange) and for the even numbers (2, 4, 6, and 8; color-coded in blue) separately. RT = response time. See the online article for the color version of this figure.

Figure D2

Overall RTs per Number Magnitude in Color Intensity Judgment (Experiment 2)



Note. Overall RTs per number magnitude averaged across all trials from all participants, with error bars representing ±1 standard error, with a regression line representing slope estimates for the odd numbers (1, 3, 7, and 9; color-coded in orange) and for the even numbers (2, 4, 6, and 8; color-coded in blue) separately. RT = response time. See the online article for the color version of this figure.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

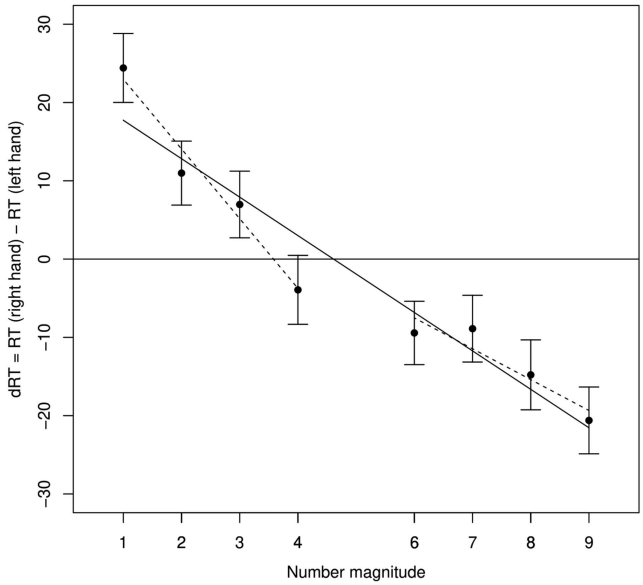
This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix E

Overall dRTs for Participants Revealing a Reliable Regular SNARC Effect

Figure E1

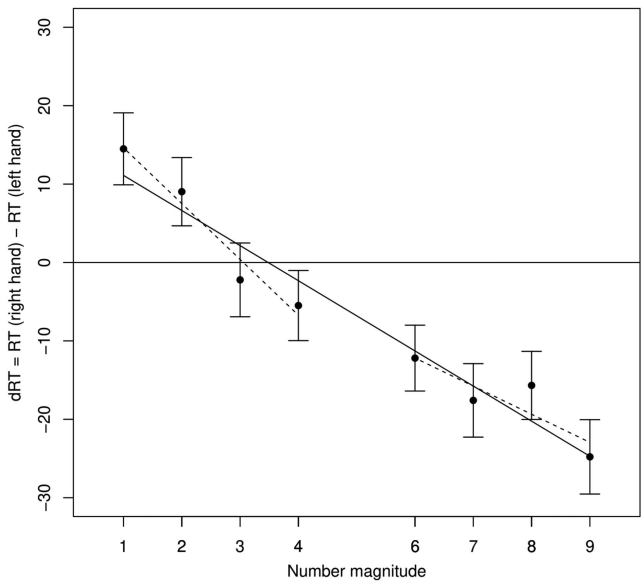
Overall dRTs per Number Magnitude for Participants With a Reliable Regular SNARC Effect in Nominal Color Judgment (Experiment 1)



Note. Overall dRTs per number magnitude averaged across all trials from the participants who revealed a reliable regular SNARC effect (48% of the sample, according to the 90% H0 bootstrap confidence intervals), with error bars representing ±1 standard error, with regression lines representing slope estimates for the whole stimulus set (solid line) as well as for the lower range (number magnitudes 1, 2, 3, and 4) and for the higher range (number magnitudes 6, 7, 8, and 9) separately (dashed lines). dRT = differences in response times between both hands; SNARC = Spatial–Numerical Association of Response Codes; RT = response time; H0 = null hypothesis.

Figure E2

Overall dRTs per Number Magnitude for Participants With a Reliable Regular SNARC Effect in Color Intensity Judgment (Experiment 2)



Note. Overall dRTs per number magnitude averaged across all trials from the participants who revealed a reliable regular SNARC (41% of the sample, according to the 90% H0 bootstrap confidence intervals), with error bars representing ±1 standard error, with regression lines representing slope estimates for the whole stimulus set (solid line) as well as for the lower range (number magnitudes 1, 2, 3, and 4) and for the higher range (number magnitudes 6, 7, 8, and 9) separately (dashed lines). dRT = differences in response times between both hands; SNARC = Spatial–Numerical Association of Response Codes; RT = response time; H0 = null hypothesis.

Appendix F

Exploratory Analysis: Linear Mixed Models

On top of repeated-measures regressions, H0 bootstrapping, and the between-experiment comparison, we ran a LMM for each experiment, in which RTs of all participants were regressed on number magnitude, number parity, response side, color, and handedness. This allowed for testing main effects of all five predictors to find RT differences between number magnitudes (i.e., size effect; Fias et al., 2003), odd versus even numbers (i.e., odd effect; Hines, 1990), left- versus right-hand responses (i.e., dominant-hand effect; Wang et al., 2022; and, since there are more right- than left-handers, right-hand responses are on average faster), cyan versus yellow or light versus dark cyan numbers (e.g., faster responses to dark than to light achromatic circles; Ren et al., 2011; faster responses to dark than to light chromatic rectangles, but also to light than to dark chromatic arrows; Wang et al., 2022), and left- versus right-handers. Note that handedness was a participant-level predictor, while all others are item-level predictors.

Apart from the main effects, we investigated several interaction effects in both experiments. First, the presence of SNARC or MARC effects was tested by the interactions of response side and number magnitude or number parity, respectively, to confirm the results from the repeated-measures regression. Second, we tested for the PNARC effect (Parity–Numerical Association of Representational Codes; Krajcsi et al., 2018) by the interaction of number magnitude and number parity. The PNARC effect describes the association of number magnitude and parity, but its direction remains unclear; some of the participants in Krajcsi et al.'s (2018) study associated oddness/evenness with small/large number magnitude, respectively, but other participants showed a reversed association. Last, in both experiments, we investigated the interaction of handedness with the SNARC and MARC effects (i.e., three-way interaction between handedness, response side, and number magnitude or number parity,

(*Appendices continue*)

respectively) to check whether they differ between right-handers and left-handers (null effects for the SNARC effect found by Cipora, Soltanlou, et al., 2019; Dehaene et al., 1993; Huber et al., 2015; regular MARC effect in right-handers and absent or even reversed MARC effect in left-handers found by Cipora, Soltanlou, et al., 2019; Huber et al., 2015). In Experiment 2, as preregistered, potential luminance associations were explored with two more interaction effects. First, we tested the interaction of color intensity and response side to compare with the results of Ren et al. (2011), Fumarola et al. (2014), and Wang et al. (2022), who observed different associations of luminance and directional space. Second, we investigated the interaction of color intensity and number magnitude to compare with results by Pinel et al. (2004), Cohen Kadosh and Henik (2006), and Cohen Kadosh, Lammertyn, and Izard (2008). Random intercepts per participant accounted for interindividual variance between participants in RTs in both LMMs. However, because of the large number of predictor terms (five main effects, five two-way interactions, two three-way interactions, random intercepts), we did not include random slopes per predictor term per participant.

After defining a start model including all potential predictors, we performed an automatic backward LMM selection procedure by eliminating nonsignificant terms ($\alpha = .05$ for both fixed and random effects) with the *step* function of the *lmerTest* R package (Kuznetsova et al., 2017). This selection procedure follows the principle of marginality, that is, highest order interactions are tested first, and lower order effects are only tested and potentially eliminated in case of nonsignificance of higher order effects including the respective predictors. For the resulting model, we ran overall *F*-tests for all *b*-coefficients, while estimating the number of degrees of freedom with Satterthwaite's method. Furthermore, we estimated the *b*-coefficients for each factor level and provide exact results and parameter estimates with standard errors in Tables F1 and F2. Effects are reported starting with the highest level interaction remaining in the resulting model.

Experiment 1: Nominal Color Judgment

The start model for the automatic backward LMM selection looked as follows:

$$\begin{aligned} \mathrm{RT}_i = {} & b_0 + b_1 \times \text{magnitude} + b_2 \times \text{parity} + b_3 \times \text{side} \\ & + b_4 \times \text{color} + b_5 \times \text{handedness} + b_6 \times \text{magnitude} \times \text{side} \\ & + b_7 \times \text{parity} \times \text{side} + b_8 \times \text{magnitude} \times \text{parity} \\ & + b_9 \times \text{magnitude} \times \text{side} \times \text{handedness} \\ & + b_{10} \times \text{parity} \times \text{side} \times \text{handedness} + v_{0i}. \end{aligned} \quad \text{(F1)}$$

The resulting model did not include the MARC effect (two-way interaction between number parity and response side, b_7) and the interaction between the MARC effect and handedness (three-way interaction between number parity, response side, and handedness, b_{10}):

$$\begin{aligned} \mathrm{RT}_i = {} & b_0 + b_1 \times \text{magnitude} + b_2 \times \text{parity} + b_3 \times \text{side} \\ & + b_4 \times \text{color} + b_5 \times \text{handedness} + b_6 \times \text{magnitude} \times \text{side} \\ & + b_8 \times \text{magnitude} \times \text{parity} \\ & + b_9 \times \text{magnitude} \times \text{side} \times \text{handedness} + v_{0i}. \end{aligned} \quad \text{(F2)}$$

Overall *F*-tests for all *b*-coefficients as well as their estimates are summarized in Table F1. The three-way interaction between magnitude, response side, and handedness (b_9) remained in the model, reflecting that the SNARC effect depends on handedness. Descriptively, SNARC slopes are most negative for right-handers and still negative but weaker for left-handers, with ambidextrous participants revealing intermediate SNARC slopes. Second, we found an interaction between magnitude and response side (b_6), reflecting the SNARC effect (over all handedness levels). Third, the data revealed an interaction between magnitude and parity (b_8), reflecting that small/large number magnitude was associated to even/odd, respectively. This is visible in Figure D1, where responses

Table F1

F-Statistics and p Values for All Fixed Effects in the LMM for Experiment 1 Remaining After the Backward LMM Selection Procedure

Effect	*F*-statistic	*p* value	Parameter	Estimate	*SE*
Intercept			b_0	439.556	4.072
Magnitude	0.769	.381	b_1	0.809	0.179
Parity	9.363	.002*	$b_{2\ \text{even}}$	−4.335	1.417
Side	11.450	<.001*	$b_{3\ \text{right}}$	4.578	1.353
Color	4.375	<.001*	$b_{4\ \text{yellow}}$	1.359	0.650
Handedness	1.098	.335	$b_{5\ \text{left-handers}}$	20.246	13.701
			$b_{5\ \text{ambidextrous}}$	−0.668	21.286
Magnitude × Side	59.068	<.001*	$b_{6\ \text{Magnitude} \times \text{Right}}$	−1.851	0.241
Magnitude × Parity	4.260	.039*	$b_{8\ \text{Magnitude} \times \text{Even}}$	0.520	0.252
Magnitude × Side × Handedness	3.636	.006*	$b_{9\ \text{Magnitude} \times \text{Left} \times \text{Left-Handers}}$	−1.015	0.480
			$b_{9\ \text{Magnitude} \times \text{Left} \times \text{Ambidextrous}}$	−0.098	0.745
			$b_{9\ \text{Magnitude} \times \text{Right} \times \text{Left-Handers}}$	0.544	0.480
			$b_{9\ \text{Magnitude} \times \text{Right} \times \text{Ambidextrous}}$	0.266	0.745
Random intercept	15,965	<.001*	v_{0i}	68.21	

Note. We provide dummy-coded parameter estimates (with response times for number magnitude 0, for odd numbers, for the right response side, for yellow color, and for right-handed participants serving as reference category) with their *SE*s. For the random intercepts per participant, instead of an *F*-statistic, a χ^2-statistic with one degree of freedom and the corresponding *p* value resulting from the Likelihood-Ratio Test as well as the standard deviation as estimate are reported. LMM = Linear Mixed Model; *SE* = standard error.
* $p < .05$, significant predictor.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

to even numbers get slower with increasing magnitude but responses to odd numbers get faster with increasing magnitude.

Regarding main effects, note that the calculation of the estimates and significance tests were made while retaining all previously reported interaction effects in the LMM. We therefore also analyzed the main effects separately and report these results in Appendix G. The LMM revealed main effects of number parity (b_2), such that responses were faster to even than to odd numbers (thus replicating the odd effect as introduced by Hines, 1990), response side (b_3), reflected by faster responses with the left than with the right hand (standing in contrast with the dominant-hand effect; Wang et al., 2022), and color (b_4), such that responses were faster to cyan than to yellow. Note that the main effects of number magnitude (b_1) and handedness (b_5) remained nonsignificant, although included in the resulting model because of the principle of marginality. Finally, we observed differences in RTs between participants reflected by significant random intercepts.

Experiment 2: Color Intensity Judgment

The start model for the automatic backward LMM selection looked as follows:

$$\begin{aligned} \mathrm{RT}_i = {} & b_0 + b_1 \times \text{magnitude} + b_2 \times \text{parity} + b_3 \times \text{side} \\ & + b_4 \times \text{color} + b_5 \times \text{handedness} \\ & + b_6 \times \text{magnitude} \times \text{side} + b_7 \times \text{parity} \times \text{side} \\ & + b_8 \times \text{magnitude} \times \text{parity} + b_9 \times \text{side} \times \text{color} \\ & + b_{10} \times \text{magnitude} \times \text{color} \\ & + b_{11} \times \text{magnitude} \times \text{side} \times \text{handedness} \\ & + b_{12} \times \text{parity} \times \text{side} \times \text{handedness} + v_{0i}. \end{aligned} \tag{F3}$$

The resulting model did not include the MARC effect (the two-way interaction between number parity and response side, b_7) or the interaction between the MARC effect and handedness (the three-way interaction between number parity, response side, and handedness, b_{12}):

$$\begin{aligned} \mathrm{RT}_i = {} & b_0 + b_1 \times \text{magnitude} + b_2 \times \text{parity} + b_3 \times \text{side} \\ & + b_4 \times \text{color} + b_5 \times \text{handedness} \\ & + b_6 \times \text{magnitude} \times \text{side} + b_8 \times \text{magnitude} \times \text{parity} \\ & + b_9 \times \text{side} \times \text{color} + b_{10} \times \text{magnitude} \times \text{color} \\ & + b_{11} \times \text{magnitude} \times \text{side} \times \text{handedness} + v_{0i}. \end{aligned} \tag{F4}$$

Overall F-tests for all b-coefficients as well as their estimates are summarized in Table F2. The three-way interaction between magnitude, response side, and handedness (b_{11}) was part of the resulting model, showing that the SNARC effect depends on handedness. Descriptively, SNARC slopes are most negative for right-handed participants and similar for ambidextrous and left-handed participants, with left-handed participants showing the weakest SNARC effect. Second, our data revealed an interaction between magnitude and parity (b_8), reflecting that small/large number magnitude was associated to even/odd, respectively. This is visible in Figure D2, where responses to even numbers get slower with increasing magnitude but responses to odd numbers get faster with increasing magnitude. Third, we observed an interaction between response side and color intensity (b_9), such that light/dark numbers were associated with the left/right respectively. Fourth, we found an interaction between number magnitude and color intensity (b_{10}), indicating that small/large numbers were associated with light/dark numbers, respectively.

Again, as for Experiment 1, main effects reported here were obtained while retaining all significant interaction effects in the LMM, which is why we report separate analyses of the main effects in Appendix G. The LMM revealed a main effect of number parity (b_2), such that responses were faster to even than to odd numbers

Table F2

F-Statistics and p Values for All Fixed Effects in the LMM for Experiment 2 Remaining After the Backward LMM Selection Procedure

Effect	F-statistic	p value	Parameter	Estimate	SE
Intercept			b_0	457.272	4.367
Magnitude	0.061	.804	b_1	1.282	0.224
Parity	5.152	.023*	$b_{2\ \text{even}}$	−3.347	1.474
Side	0.488	.485	$b_{3\ \text{right}}$	3.071	1.562
Color	17.956	<.001*	$b_{4\ \text{dark}}$	8.055	1.562
Handedness	1.501	.224	$b_{5\ \text{left-handers}}$	−21.094	13.880
			$b_{5\ \text{ambidextrous}}$	15.895	21.510
Magnitude × Side	0.073	.786	$b_{6\ \text{Magnitude} \times \text{Right}}$	−0.128	0.251
Magnitude × Parity	8.322	.004*	$b_{8\ \text{Magnitude} \times \text{Even}}$	0.756	0.262
Side × Color	9.526	.002*	$b_{9\ \text{Right} \times \text{Dark}}$	−4.176	1.353
Magnitude × Color	24.033	<.001*	$b_{10\ \text{Magnitude} \times \text{Dark}}$	−1.211	0.247
Magnitude × Side × Handedness	3.018	.017*	$b_{11\ \text{Magnitude} \times \text{Left} \times \text{Left-Handers}}$	−0.313	0.480
			$b_{11\ \text{Magnitude} \times \text{Left} \times \text{Ambidextrous}}$	−0.884	0.744
			$b_{11\ \text{Magnitude} \times \text{Right} \times \text{Left-Handers}}$	0.929	0.480
			$b_{11\ \text{Magnitude} \times \text{Right} \times \text{Ambidextrous}}$	0.218	0.744
Random intercept	14,834	<.001*	v_{0i}	68.83	

Note. We provide dummy-coded parameter estimates (with response times for number magnitude 0, for odd numbers, for the right response side, for light color intensity, and for right-handed participants serving as reference category) with their *SE*s. For the random intercepts per participant, instead of an F-statistic, a χ^2-statistic with one degree of freedom and the corresponding p value resulting from the Likelihood-Ratio Test as well as the standard deviation as estimate are reported. LMM = Linear Mixed Model; *SE* = standard error.

* $p < .05$, significant predictor.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

(thus replicating the odd effect as introduced by Hines, 1990), and a main effect of color intensity (b_4), such that responses were faster to light than to dark cyan. Note that the main effects of number magnitude (b_1), response side (b_3), and handedness (b_5), as well as the SNARC effect (two-way interaction of magnitude and response side, b_6), all remained nonsignificant, although included in the resulting model because of the principle of marginality. Finally, we observed differences in RTs between participants reflected by significant random intercepts.

Both experiments revealed an association of small/large numbers with even/odd, respectively, reflecting the PNARC effect (Krajcsi et al., 2018; Nuerk et al., 2004). Krajcsi et al. (2018) give an overview of theories according to which parity and magnitude should or should not be related. According to the Polarity Correspondence Principle (Proctor & Cho, 2006) and to the Linguistic Markedness account (Nuerk et al., 2004), which both might be plausible explanations for the PNARC effect, some stimulus features (e.g., for numerical stimuli: small or odd) and response sides (e.g., the left) are coded with a negative polarity or linguistically marked, whereas other stimulus features (e.g., for numerical stimuli: large or even) and response sides (e.g., the right) are coded with a positive polarity or linguistically unmarked. Hence, the SNARC, MARC, and PNARC effects might result from the polarity or markedness correspondence between the three dimensions of number magnitude, number parity, and space. However, from this relationship between the three effects and the three dimensions (as illustrated in Figure 3 in Krajcsi et al., 2018), one would predict an association of small/large numbers with odd/even, respectively, which is opposite to the pattern we found. A potential explanation for our findings is that the relationship between magnitude and parity might be different in color judgment tasks, where the SNARC and MARC effects seem to be driven only or primarily by small numbers.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix G

Exploratory Analysis: Effects on RTs

On top of preregistered analyses, we tested potential influences on RTs, some of them reflecting typical numerical-cognition effects. Specifically, we investigated the effects of number magnitude (replication of the size effect; Fias et al., 2003), number parity (replication of the odd effect; Hines, 1990), response side (replication of the dominant-hand effect in combination with the majority of the sample being right-handed; Wang et al., 2022), color, and handedness on RTs.

Experiment 1: Nominal Color Judgment

When regressing RTs on magnitude for each participant separately, our two-sided one-sample t test of the obtained slopes against zero revealed that responses got slightly faster with increasing number magnitude, $\text{slope}_{\text{size}} = -0.22$ ms, $t(339) = -2.31$, $p = .021$, $BF_{10} = 0.84$, which would stand in contrast with the typically observed size effect (Fias et al., 2003). However, Bayesian evidence was inconclusive, and visual inspection of Figure D1 shows that this is only due to comparatively slow reactions to Number 1, while there is almost no difference in RTs between Numbers 2 and 9.

To test the effects of parity, response side, and color, we compared responses for odd and even numbers, with the left and right response keys, and for cyan and yellow stimuli in two-sided paired t tests. Responses were significantly slower to odd than to even numbers, $\text{estimate}_{\text{odd}} = 1.74$ ms, $t(339) = 3.26$, $p = .001$, $BF_{10} = 10.99$, replicating the odd effect (Hines, 1990) with strong Bayesian evidence (see Figure D1). The effect of response side was significant, $\text{estimate}_{\text{side}} = 3.98$ ms, $t(339) = 4.26$, $p < .001$, $BF_{10} = 382.93$, indicating faster responses with the right side. The strong Bayesian evidence for this result is not surprising in a mostly right-handed sample. The effect of color was not significant, $\text{estimate}_{\text{color}} = -1.36$ ms, $t(339) = -1.78$, $p = .076$, $BF_{10} = 0.29$, with moderate Bayesian evidence against it. Average RTs depending on magnitude, parity, side, and color can be seen in Figure D1.

Lastly, mean RTs were compared between the 301 right-handed, the 28 left-handed, and the 11 ambidextrous participants with a one-way analysis of variance (ANOVA), which revealed no significant difference between the three subsamples, $F(2, 337) = 0.99$, $p = .373$, $BF_{10} = 0.23$, supported by moderate Bayesian evidence against a difference.

Experiment 2: Color Intensity Judgment

In line with our preregistration, we already tested potential effects of number magnitude, number parity, response side, color intensity, and handedness on RTs within the LMM. However, as described above, some interactions turned out to be significant and were therefore retained in the final LMM, so that the calculation of estimates and significance tests of the main effects takes these into account. Hence, we additionally calculated the effects on RTs separately, as typically reported in the literature and as done for Experiment 1.

When regressing RTs on magnitude for each participant separately, our two-sided one-sample t test did not reveal that the obtained slopes differed significantly from zero, $\text{slope}_{\text{size}} = -0.08$ ms, $t(311) = -0.70$, $p = .486$, $BF_{10} = 0.08$, with the Bayes Factor providing strong evidence against the size effect (Fias et al., 2003).

To test the effects of parity, response side, and color, we compared responses for odd and even numbers, with the left and right response keys, and for dark and light cyan stimuli in two-sided paired t tests. The effect of parity was not significant, $\text{estimate}_{\text{odd}} = -0.43$ ms, $t(311) = -0.78$, $p = .433$, $BF_{10} = 0.09$, with strong Bayesian evidence against the odd effect (Hines, 1990). The effect of response side was significant, $\text{estimate}_{\text{side}} = 4.68$ ms, $t(311) = 4.43$, $p < .001$, $BF_{10} = 793.07$, with strong Bayesian evidence, indicating faster responses with the right than with the left side. The effect of color intensity was not significant, $\text{estimate}_{\text{color}} = 0.09$ ms, $t(311) = 0.10$, $p = .919$, $BF_{10} = 0.06$, and the Bayes Factor indicated strong evidence against a difference between the RTs for dark and light cyan stimuli. Average

(*Appendices continue*)

RTs depending on magnitude, parity, side, and color intensity can be seen in Figure D2.

Last, we conducted a one-way ANOVA to compare mean RTs between the 273 right-handed, the 28 left-handed, and the 11 ambidextrous participants, which revealed no significant difference between the three subsamples, $F(2, 309) = 1.29$, $p = .277$, $BF_{10} = 0.27$, with moderate Bayesian evidence against a difference.

Appendix H

Exploratory Analysis: Influences of Participant Characteristics

Potential influences of participant characteristics on the SNARC and MARC effects were investigated in nonpreregistered analyses. Specifically, we looked at possible effects of synesthesia, handedness, math skills, gender, first language, and finger-counting habits and the stability of these habits.

Experiment 1: Nominal Color Judgment

Synesthesia

First, we tested the SNARC and MARC effects in the 30 participants who seem to be affected by synesthesia (i.e., who answered the question whether numbers between 1 and 9 are generally represented in certain colors in their mind with "yes" and additionally left a comment that points toward synesthesia).[H1] The repeated-measures regression revealed no significant SNARC, $slope_{SNARC} = -1.54$ ms, $SD = 6.46$, $t(29) = -1.31$, $p = .202$, $BF_{10} = 0.42$, or St-SNARC effect, $slope_{St\text{-}SNARC} = -0.07$, $SD = 0.28$, $t(29) = -1.45$, $p = .158$, $BF_{10} = 0.50$, on the group level, but Bayesian evidence remained inconclusive. Moreover, we found moderate Bayesian evidence against a MARC, $slope_{MARC} = -0.30$ ms, $SD = 15.05$, $t(29) = -0.11$, $p = .913$, $BF_{10} = 0.20$, or St-MARC effect, $slope_{St\text{-}MARC} = -0.01$, $SD = 0.18$, $t(29) = -0.33$, $p = .742$, $BF_{10} = 0.20$, in the synesthesia subsample. However, the difference between the synesthesia subsample (30 participants) and the rest of our sample (310 participants) was not significant for any of the effects, $t_{SNARC}(31.21) = -0.16$, $p = .876$, $BF_{10} = 0.21$; $t_{St\text{-}SNARC}(33.25) = -0.75$, $p = .461$, $BF_{10} = 0.28$; $t_{MARC}(40.52) = -0.82$, $p = .418$, $BF_{10} = 0.24$; $t_{St\text{-}MARC}(38.73) = -0.86$, $p = .396$, $BF_{10} = 0.25$, with moderate Bayesian evidence against a difference. Note that the power for small effects is probably too low here, because the synesthesia subsample only consisted of 30 participants. We therefore investigated by means of bootstrapping with 5,000 iterations how frequently the effects would be significant in nonsynesthesia subsamples of the same size (i.e., with 30 slopes) that were randomly drawn from all 310 nonsynesthesia participants. Significant effects would have arisen in 64% for SNARC, 72% for St-SNARC, 13% for MARC, and 16% for St-MARC.

Handedness

The 28 left-handers revealed no significant SNARC, $slope_{SNARC} = -0.90$ ms, $SD = 4.76$, $t(27) = -1.00$, $p = .328$, $BF_{10} = 0.31$, or St-SNARC effect, $slope_{St\text{-}SNARC} = -0.11$, $SD = 0.28$, $t(27) = -1.98$, $p = .058$, $BF_{10} = 1.10$. Moreover, left-handed participants did not show a MARC, $slope_{MARC} = 2.48$ ms, $SD = 27.53$, $t(27) = 0.48$, $p = .638$, $BF_{10} = 0.22$, or St-MARC effect, $slope_{St\text{-}MARC} = -0.05$, $SD = 0.27$, $t(27) = -1.01$, $p = .321$, $BF_{10} = 0.32$. However, the difference between the left-handed (28 participants) and the right-handed (301 participants) subsamples did not reach significance for any of the effects, $t_{SNARC}(31.11) = 0.89$, $p = .378$, $BF_{10} = 0.32$; $t_{St\text{-}SNARC}(30.67) = 0.01$, $p = .993$, $BF_{10} = 0.21$; $t_{MARC}(29.52) = 1.04$, $p = .305$, $BF_{10} = 0.50$; $t_{St\text{-}MARC}(30.53) = -0.23$, $p = .819$, $BF_{10} = 0.22$. In all these tests, Bayesian evidence remained inconclusive or was only moderately in favor of the null hypothesis. Again, we investigated by means of bootstrapping with 5,000 iterations how frequently the effects would be significant in right-handed subsamples of the same size (i.e., with 28 slopes) that were randomly drawn from all 301 right-handed participants. Significant effects would have arisen in 57% for SNARC, 62% for St-SNARC, 16% for MARC, and 14% for St-MARC.

Math Skills

Next, we investigated whether there was a relationship between self-estimated math skills and unstandardized SNARC and MARC effects. Pearson's correlation coefficient was not significantly different from 0 for any effect, $r_{SNARC} = 0.042$, $t(338) = 0.78$, $p = .438$, $BF = 0.17$; $r_{MARC} = -0.062$, $t(338) = -1.15$, $p = .251$, $BF_{10} = 0.24$, and Bayes Factors revealed moderate evidence against both correlations.

Gender

The 134 women revealed a significant SNARC effect, $slope_{SNARC} = -1.34$ ms, $SD = 4.42$, $t(133) = -3.52$, $p < .001$, $BF_{10} = 31.62$, and St-SNARC effect, $slope_{St\text{-}SNARC} = -0.08$, $SD = 0.24$, $t(133) = -3.84$, $p < .001$, $BF_{10} = 92.12$. Moreover, women seemed not to show a MARC effect, $slope_{MARC} = -1.54$ ms, $SD = 22.17$, $t(133) = -0.80$, $p = .424$, $BF_{10} = 0.13$, or St-MARC effect, $slope_{St\text{-}MARC} = -0.03$, $SD = 0.23$, $t(133) = -1.66$, $p = .100$, $BF_{10} = 0.36$. The 202 men revealed a significant SNARC effect, $slope_{SNARC} = -1.93$ ms, $SD = 4.20$, $t(201) = -6.55$, $p < .001$, $BF_{10} = 16{,}949{,}183.39$, and St-SNARC effect, $slope_{St\text{-}SNARC} = -0.13$, $SD = 0.24$, $t(201) = -7.44$, $p < .001$, $BF_{10} = 2{,}338{,}345{,}439.94$. Interestingly, men seemed to reveal a MARC effect, $slope_{MARC} = -3.26$ ms, $SD = 19.14$ $t(201) = -2.42$, $p = .016$, $BF_{10} = 1.36$, and St-MARC effect, $slope_{St\text{-}MARC} = -0.04$, $SD = 0.23$, $t(201) = -2.77$, $p = .006$, $BF_{10} = 3.18$. However, the difference between women and men did not reach significance for any of the effects, $t_{SNARC}(247.47) = 1.22$, $p = .223$, $BF_{10} = 0.25$;

[H1] Importantly, according to the synesthesia assessment, the synesthesia subsamples in Experiments 1 and 2 consisted of 30 and 40 participants, respectively. When putting these 70 participants in relation to the total sample size of 652 participants, the resulting proportion of 11% is very high as compared to the literature claiming grapheme-color synesthesia to affect 1% (Simner et al., 2006) or even a smaller proportion of the population (Rich et al., 2005). Our synesthesia assessment might have a high sensitivity, but a low specificity. In other words, some participants might have been wrongly classified as affected by synesthesia. Therefore, in case a true difference between participants who are affected by synesthesia and those who are not exists in the SNARC and MARC effects, our synesthesia subsample might therefore not have been typical enough to detect it.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

$t_{\text{St-SNARC}}(283{,}59) = 1.69$, $p = .093$, $BF_{10} = 0.25$; $t_{\text{MARC}}(255.65) = 0.74$, $p = .462$, $BF_{10} = 0.16$; $t_{\text{St-MARC}}(286.84) = -0.48$, $p = .634$, $BF_{10} = 0.14$. All four comparisons revealed moderate Bayesian evidence against a difference between the two genders.

First Language

Strong Bayesian evidence was obtained against a difference in unstandardized SNARC or MARC effects between the six languages in a one-way ANOVA, $F_{\text{SNARC}}(5, 334) = 0.88$, $p = .495$, $BF_{10} = 0.02$; $F_{\text{MARC}}(5, 334) = 1.38$, $p = .232$, $BF_{10} = 0.06$.

Finger-Counting Habits and Stability

First, we tested whether the unstandardized SNARC effect differed between the 133 participants who start finger-counting with the left hand and the 146 participants who start with the right hand in an independent-samples t test. The $\text{slope}_{\text{SNARC}} = -2.04$ ms for left-starters and $\text{slope}_{\text{SNARC}} = -1.67$ ms for right-starters did not differ significantly, $t_{\text{SNARC}}(252.68) = -0.69$, $p = .493$, $BF_{10} = 0.17$, and Bayesian evidence was moderately against a difference. Second, to investigate the relationship between the unstandardized SNARC effect and stability of these finger-counting habits, a Jonckheere-Terpstra test with 5,000 permutations was used (as in Hohol et al., 2022). We display the distributions of self-reported finger-counting habits and stability in Table 2. There was no monotone trend within left-starters ($p = .321$) or within right-starters ($p = .782$).

Experiment 2: Color Intensity Judgment

Synesthesia

First, we tested the SNARC and MARC effects in the 40 participants who seem to be affected by synesthesia.[H2] In the repeated-measures regression, both SNARC, $\text{slope}_{\text{SNARC}} = -1.04$ ms, $SD = 3.37$, $t(39) = -1.95$, $p = .059$, $BF_{10} = 0.94$, and St-SNARC, $\text{slope}_{\text{St-SNARC}} = -0.07$, $SD = 0.23$, $t(39) = -1.83$, $p = .074$, $BF_{10} = 0.78$, failed to reach significance, and the evidence remained inconclusive according to the Bayes Factors. The synesthesia subsample did not show a MARC either, $\text{slope}_{\text{MARC}} = 1.64$ ms, $SD = 23.76$, $t(39) = 0.44$, $p = .665$, $BF_{10} = 0.19$, or St-MARC, $\text{slope}_{\text{St-MARC}} = 0.03$, $SD = 0.25$, $t(39) = 0.69$, $p = .495$, $BF_{10} = 0.21$, but here Bayes Factors provided moderate evidence against a MARC effect. However, the difference between the synesthesia subsample (40 participants) and the rest of our sample (272 participants) was not significant for any effect, and we observed moderate Bayesian evidence against a difference, $t_{\text{SNARC}}(56.57) = -0.19$, $p = .852$, $BF_{10} = 0.18$; $t_{\text{St-SNARC}}(54.78) = -0.39$, $p = .695$, $BF_{10} = 0.19$; $t_{\text{MARC}}(46.28) = -0.41$, $p = .684$, $BF_{10} = 0.20$; $t_{\text{St-MARC}}(49.01) = -0.82$, $p = .413$, $BF_{10} = 0.26$. As for Experiment 1, the power for small effects might be too low here, so we investigated by means of bootstrapping with 5,000 iterations how frequently the effects would be significant in nonsynesthesia subsamples of the same size (i.e., with 40 slopes) that we randomly drew from all 272 nonsynesthesia participants. Significant effects would have arisen in 41% for SNARC, 50% for St-SNARC, 3% for MARC, and 4% for St-MARC.

Handedness

The 28 left-handers revealed no significant SNARC, $\text{slope}_{\text{SNARC}} = -0.08$ ms, $SD = 3.71$, $t(27) = -0.11$, $p = .909$, $BF_{10} = 0.20$, or St-SNARC effect, $\text{slope}_{\text{St-SNARC}} = -0.01$, $SD = 0.26$, $t(27) = -0.11$, $p = .910$, $BF_{10} = 0.20$, and Bayesian evidence was moderate against these effects. Moreover, left-handed participants did not show a MARC, $\text{slope}_{\text{MARC}} = -5.95$ ms, $SD = 17.18$, $t(27) = -1.83$, $p = .078$, $BF_{10} = 0.87$, or St-MARC effect, $\text{slope}_{\text{St-MARC}} = -0.06$, $SD = 0.22$, $t(27) = -1.51$, $p = .143$, $BF_{10} = 0.55$, but Bayesian evidence remained inconclusive. However, the difference between the left-handed (28 participants) and the right-handed (273 participants) subsamples did not reach significance for any of the effects, $t_{\text{SNARC}}(33.58) = 1.62$, $p = .114$, $BF_{10} = 0.60$; $t_{\text{St-SNARC}}(32.33) = 1.65$, $p = .108$, $BF_{10} = 0.76$; $t_{\text{MARC}}(34.40) = -1.92$, $p = .063$, $BF_{10} = 0.81$; $t_{\text{St-MARC}}(33.23) = -1.42$, $p = .162$, $BF_{10} = 0.49$, but Bayesian evidence again remained inconclusive. Again, we investigated by means of bootstrapping with 5,000 iterations how frequently the effects would be significant in right-handed subsamples of the same size (i.e., with 28 slopes) that were randomly drawn from all 273 right-handed participants. Significant effects would have arisen in 38% for SNARC, 45% for St-SNARC, 4% for MARC, and 4% for St-MARC.

Math Skills

Next, we investigated the potential relationship between self-estimated math skills and unstandardized SNARC and MARC effects. Pearson's correlation coefficient was not significantly different from 0 for the effects, $r_{\text{SNARC}} = -0.058$, $t(310) = -1.02$, $p = .311$, $BF_{10} = 0.22$; $r_{\text{MARC}} = 0.054$, $t(310) = 0.95$, $p = .342$, $BF_{10} = 0.21$, with Bayes Factors speaking moderately against both correlations.

Gender

The 134 women revealed a significant SNARC effect, $\text{slope}_{\text{SNARC}} = -0.99$ ms, $SD = 3.368$, $t(133) = -3.11$, $p = .002$, $BF_{10} = 9.27$, and St-SNARC effect, $\text{slope}_{\text{St-SNARC}} = -0.07$, $SD = 0.24$, $t(133) = -3.26$, $p = .001$, $BF_{10} = 14.40$. Women seemed not to show a MARC effect, $\text{slope}_{\text{MARC}} = -0.06$ ms, $SD = 20.61$, $t(133) = -0.03$, $p = .973$, $BF_{10} = 0.10$, or St-MARC effect, $\text{slope}_{\text{St-MARC}} = -0.01$, $SD = 0.23$, $t(133) = -0.45$, $p = .652$, $BF_{10} = 0.11$. The 172 men revealed a significant SNARC effect, $\text{slope}_{\text{SNARC}} = -1.28$ ms, $SD = 4.11$, $t(171) = -4.09$, $p < .001$, $BF_{10} = 219.46$, and St-SNARC effect, $\text{slope}_{\text{St-SNARC}} = -0.09$, $SD = 0.26$, $t(171) = -4.49$, $p < .001$, $BF_{10} = 1003.10$. Men seemed not to reveal a MARC effect, $\text{slope}_{\text{MARC}} = -0.26$ ms, $SD = 17.96$ $t(171) = -0.19$, $p = .847$, $BF_{10} = 0.09$, and St-MARC effect, $\text{slope}_{\text{St-MARC}} < 0.01$, $SD = 0.22$, $t(171) = -0.06$, $p = .956$, $BF_{10} = 0.09$. There seemed not to be any difference between women and men for any of the effects, $t_{\text{SNARC}}(298.05) = 0.66$, $p = .513$, $BF_{10} = 0.15$; $t_{\text{St-SNARC}}(298.42) = 0.84$, $p = 402$, $BF_{10} = 0.18$; $t_{\text{MARC}}(264.78) = -0.14$, $p = .885$, $BF_{10} = 0.13$; $t_{\text{St-MARC}}(280.65) = -0.38$, $p = .704$, $BF_{10} = 0.14$. All four comparisons revealed moderate Bayesian evidence against a difference between the two genders.

[H2] See Footnote H1 for synesthesia prevalence and assessment in both experiments.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

First Language

No significant difference between the six languages and strong Bayesian evidence against a difference were found for unstandardized SNARC or MARC effects in one-way ANOVAs, $F_{SNARC}(5, 306) = 0.13$, $p = .985$, $BF_{10} = 0.01$; $F_{MARC}(5, 306) = 0.70$, $p = .624$, $BF_{10} = 0.02$.

Finger-Counting Habits and Stability

First, we tested whether the unstandardized SNARC effect differed between the 87 participants who start finger-counting with the left hand and the 157 participants who start with the right hand in an independent-samples *t* test. The descriptively observed difference between $slope_{SNARC} = -1.30$ ms for left-starters and $slope_{SNARC} = -1.06$ ms for right-starters did not reach significance, with moderate Bayesian evidence against it, $t_{SNARC}(148.14) = -0.41$, $p = .682$, $BF_{10} = 0.16$. Second, to investigate the relationship between the unstandardized SNARC effect and stability of these finger-counting habits, we run a Jonckheere-Terpstra test with 5,000 permutations (as in Hohol et al., 2022). We display the distributions of self-reported finger-counting habits and stability in Table 2. There was no monotone trend within left-starters ($p = .370$) or within right-starters ($p = .743$).

Appendix I

Exploratory Analysis: Differences and Correlations Between Blocks

In this nonpreregistered analysis, we investigated whether the SNARC and MARC effects differed between the first and second block that participants completed. Importantly, in color judgment tasks, SNARC-compatible and -incompatible as well as MARC-compatible and -incompatible trials are intermixed within each experimental block. In case there is a spatial mapping of cyan or yellow (Experiment 1) or of light or dark (Experiment 2) on the left or right, there would be one block that is compatible with this and another one that is incompatible with this. However, because we do not know about such compatibility effects, we therefore only tested whether there was a difference in unstandardized SNARC and MARC effects between the first and second block that participants completed.

Experiment 1: Nominal Color Judgment

The paired *t* test did not reveal any difference and we found moderate Bayesian evidence against differences, $t_{SNARC}(339) = -0.67$, $p = .506$, $BF_{10} = 0.08$; $t_{MARC}(339) = 0.86$, $p = .393$, $BF_{10} = 0.09$. Pearson's correlation between the first and the second block was not significantly different from 0 for both effects, $r_{SNARC} = 0.054$, $t(338) = 0.99$, $p = .321$, $BF_{10} = 0.21$; $r_{MARC} = 0.074$, $t(338) = 1.36$, $p = .175$, $BF_{10} = 0.31$, with moderate evidence against the correlations.

Experiment 2: Color Intensity Judgment

The paired *t* test revealed that the unstandardized SNARC effect was significantly stronger in the first than in the second block, $t_{SNARC}(311) = -2.44$, $p = .015$, $BF_{10} = 1.18$, although Bayesian evidence remained inconclusive. There was no significant difference between blocks in the MARC effect and moderate Bayesian evidence against a difference, $t_{MARC}(311) = 0.62$, $p = .533$, $BF_{10} = 0.08$. Pearson's correlation between the first and the second block was not significantly different from 0 for both effects, $r_{SNARC} = 0.053$, $t(310) = 0.93$, $p = .353$, $BF_{10} = 0.20$; $r_{MARC} = -0.064$, $t(310) = -1.12$, $p = .263$, $BF_{10} = 0.24$, with moderate Bayesian evidence against the correlations.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix J

Exploratory Analysis: Range Dependency of the SNARC Effect

On top of the regression of dRT on number magnitude performed for each participant separately, we plotted average dRTs across all trials from all participants per number magnitude. The resulting plots are displayed in Figure 1 for Experiment 1 and in Figure 4 for Experiment 2 and suggest that the group-level SNARC effects were driven only by the dRT pattern within Numbers 1, 2, and possibly 3. In contrast, dRTs for higher numbers did not seem to contribute to a group-level SNARC effect in both experiments. Therefore, we analyzed the unstandardized SNARC effect separately for the lower and higher number ranges.

Experiment 1: Nominal Color Judgment

The group-level effect was $\text{slope}_{\text{SNARC}} = -4.84$ ms in the lower range, and only $\text{slope}_{\text{SNARC}} = -0.48$ ms in the higher range, with a paired t test revealing a significant difference and strong Bayesian evidence, $t_{\text{SNARC}}(339) = -3.87$, $p < .001$, $BF_{10} = 86.40$. Additionally, we compared the unstandardized SNARC effect between ranges including only the slopes of participants showing a reliable unstandardized SNARC effect according to the 90% H0 bootstrap confidence intervals (48% of the sample). With estimates of $\text{slope}_{\text{SNARC}} = -8.09$ ms for the lower and $\text{slope}_{\text{SNARC}} = -4.27$ ms for the higher range, the paired t test revealed a significant difference, $t_{\text{SNARC}}(162) = -2.19$, $p = .030$, $BF_{10} = 0.88$, which is illustrated in Figure E1.

Experiment 2: Color Intensity Judgment

Similar to Experiment 1, we observed a group-level effect of $\text{slope}_{\text{SNARC}} = -4.40$ ms for the lower range from 1 to 4, but only $\text{slope}_{\text{SNARC}} = -0.27$ ms for the higher range from 6 to 9. Again, the paired t test revealed a significant difference, $t_{\text{SNARC}}(311) = -3.92$, $p < .001$, $BF_{10} = 108.44$, with strong Bayesian evidence for this difference. Likewise, when only looking at participants showing a reliable unstandardized SNARC effect (41% of the sample), the paired t test revealed a significant difference, $t_{\text{SNARC}}(128) = -3.27$, $p = .001$, $BF_{10} = 14.75$ between the estimated $\text{slope}_{\text{SNARC}} = -7.81$ ms for the lower and $\text{slope}_{\text{SNARC}} = -2.67$ ms for the higher range with moderate Bayesian evidence for a difference (see Figure E2).

Received April 9, 2024
Revision received August 22, 2024
Accepted October 4, 2024 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

A1.2. Corrigendum for Study 1

Roth, L., Caffier, J., Reips, U.-D., Cipora, K., Braun, L., & Nuerk, H.-C. (in preparation). *Corrigendum to "True colors SNARC: Semantic number processing is highly automatic"*

This corrigendum refers to the following publication (see Appendix A1.1.):

Roth, L., Caffier, J., Reips, U.-D., Cipora, K., Braun, L., & Nuerk, H.-C. (2025). True colors SNARC: Semantic number processing is highly automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 51*(12), 1875–1908. https://doi.org/10.1037/xlm0001431

Corrigendum to:

"True colors SNARC: Semantic number processing is highly automatic"

Lilly Roth[1], John Caffier[2], Ulf-Dietrich Reips[2], Krzysztof Cipora[3], Lydia Braun[1], Hans-Christoph Nuerk[1,4,5*]

[1]Department of Psychology, University of Tübingen, Germany
[2]Department of Psychology, University of Konstanz, Germany
[3]Centre for Mathematical Cognition, Loughborough University, United Kingdom
[4]LEAD Graduate School & Research Network, University of Tübingen, Germany
[5]German Center for Mental Health (DZPG)

*Corresponding author:
hc.nuerk@uni-tuebingen.de

In the article “True colors SNARCing: Semantic number processing is highly automatic” published in the *Journal of Experimental Psychology: Learning, Memory, and Cognition* (https://doi.org/10.1037/xlm0001431), we investigated the automaticity of semantic number processing in two color judgment experiments. We stated that response time (RT) trimming was, after exclusions based on fixed participant-independent criteria (i.e., faster than 200 ms or slower than 1500 ms), conducted in an iterative way based on participant-dependent criteria. Specifically, we described that we computed the individual mean and standard deviation of RT separately for each participant, and excluded trials with RT falling outside the range of ±3 standard deviations from the mean. We preregistered and reported in the published article that we repeated this procedure iteratively by recomputing the mean and standard deviation after each trimming step until no further trials were excluded.

However, we would like to clarify that this individual RT trimming procedure, excluding trials with RT ±3 standard deviations from the mean, was not iteratively applied, but instead only once per participant. In fact, such single-run approaches are common practice in the field of numerical cognition. We apologize for the oversight.

Importantly, we re-ran all analyses, iteratively applying the individual RT trimming procedure, and the main results did not substantially differ from those obtained with a single-run approach. To ensure transparency, we uploaded all resulting outcomes including updated plots on the OSF repository in our folders for the project (https://osf.io/jv4as/ for Experiment 1 and https://osf.io/mwqh8/ for Experiment 2). The iterative procedure is marginally more conservative than the single-run approach. Therefore, a little more data was excluded at the trial level, and subsequently at the participant level due to more participants falling below the threshold of 75% valid trials. Note that the difference is very small, so that data from 337 instead of 340 participants remained for the analysis in Experiment 1, and data from 310 instead of 312 participants were included in Experiment 2. Crucially, these additional exclusions left the main findings qualitatively unchanged, and differences were only numerical. The group-level SNARC effect remained significant in Experiments 1 and 2, as supported by Bayesian evidence. Similarly, the MARC effect remained significant in Experiment 1, with mixed Bayesian evidence, and it remained non-significant in Experiment 2, supported by Bayesian evidence. Hence, the interpretation and discussion of the findings in the original manuscript remain valid, even when iteratively applying the individual RT trimming procedure.

A2. Study 2

A2.1. Manuscript for Study 2

Roth, L., Cipora, K., Overlander, A. T., Nuerk, H.-C., & Reips, U.-D. (under review). *Shape of SNARC: How task-dependent are Spatial-Numerical Associations? A highly powered online experiment* [manuscript for stage-2 Registered Report, after in-principle acceptance of stage-1 Registered Report by PCI-RR]. https://osf.io/jyrf3/files/8dac2

Shape of SNARC: How task-dependent are Spatial-Numerical Associations? A highly powered online experiment

Lilly Roth[1], Krzysztof Cipora[2], Annika Tave Overlander[3], Hans-Christoph Nuerk[1,4,5], Ulf-Dietrich Reips[3*]

[1]Department of Psychology, University of Tübingen, Germany

[2]Centre for Mathematical Cognition, Loughborough University, United Kingdom

[3]Department of Psychology, University of Konstanz, Germany

[4]LEAD Graduate School & Research Network, University of Tübingen, Germany

[5]German Center for Mental Health (DZPG)

*Corresponding author:

reips@uni-konstanz.de

Abstract

Spatial-Numerical Associations (SNAs) are fundamental to numerical cognition. The Spatial-Numerical Association of Response Codes (SNARC) effect denotes faster responses to small/large numbers with the left/right, respectively (Dehaene et al., 1993). It is typically measured in *magnitude classification* (MC) or *parity judgment* (PJ), where participants decide whether numbers 1–9 (excluding 5) are smaller vs. larger than 5 or odd vs. even. These tasks differ in magnitude and parity processing, compatibility effects, and other aspects. In this highly powered online experiment, data from 609 participants (Bayesian optional recruitment stopping criterion) collected in a within-subjects design demonstrated a rather categorical MC-SNARC (i.e., same left-hand advantage for 1–4, and same right-hand advantage for 6–9) also entailing continuous parts, and a continuous PJ-SNARC (i.e., right-hand advantage increasing with magnitude). These different shapes imply that the PJ-SNARC can be modeled using standard linear regression, while the MC-SNARC requires a categorical or even mixed analysis. Explained variance and split-half reliability were higher for the MC-SNARC than PJ-SNARC. We replicated the Numerical Size Effect (slower responses for larger numbers; Brysbaert, 1995), the Numerical Distance Effect (slower responses for larger numerical distance between stimulus and reference; Moyer & Landauer, 1967), and the Odd effect only in PJ (faster responses to even numbers; Hines, 1990; potentially masked by the Numerical Distance Effect in MC), but no MARC effect (faster responses to odd/even with the left/right, respectively; Nuerk et al., 2004). The MC-SNARC and PJ-SNARC correlated slightly positively, suggesting partly common underlying magnitude representations, yet being largely task-dependent.

Keywords: spatial-numerical associations, SNARC effect, magnitude classification, parity judgment, task dependency

Stage-1 Registered Report

Shape of SNARC: How task-dependent are Spatial-Numerical Associations? A highly powered online experiment

Spatial-numerical associations (SNAs) belong to the fundaments of numerical cognition (Fischer & Shaki, 2014; Toomarian & Hubbard, 2018). They have been implicated as an important representation (e.g., Dehaene et al., 2003) and as means to foster numerical and arithmetic learning (Booth et al., 2008; Dackermann et al., 2017, for an overview of embodied spatial-numerical learning). SNAs can be divided into spatial-extensional SNAs, where a particular number or magnitude is related to a physical extension (i.e., larger number to larger extensions), and directional SNAs, where a particular number is associated with a particular location in space (Patro et al., 2014). Both SNAs are important and seem to be highly dependent on the experimental situation or task (Cipora, Patro & Nuerk, 2018). Understanding such situational dependencies is key to understanding SNAs and their relation to mathematics as such (Cipora, He, & Nuerk, 2020).

SNAs can refer to explicit or implicit associations of different characteristics of numbers (e.g., cardinality, ordinality, parity) with different aspects of space, namely directions or extensions (Cipora, Haman, et al., 2020; Cipora, Schroeder, et al., 2018; Patro et al., 2014). For instance, the MARC effect (Linguistic Markedness of Response Codes; Nuerk et al., 2004) reflects the association between parity (odd/even numbers) and direction (the left/right side), respectively. The hallmark directional SNA, however, is the Spatial-Numerical Association of Response Codes (SNARC) effect, which denotes that – at least in left-to-right reading cultures – participants respond faster to small/large magnitude numbers on the left/right side, respectively (Dehaene et al., 1993). SNAs are claimed to reflect implicit and explicit mental representations of

numbers and processes operating on them (Cipora, Haman, et al., 2020). The tendency to map characteristics of numbers onto space is considered one of the basic traits of human cognition (Cipora, Patro, & Nuerk, 2018). The SNARC effect has been replicated with stimuli in different modalities and notations (e.g., visual Arabic numerals, visual number words, auditory number words, visual dice patterns; Nuerk et al., 2004; Nuerk, Wood, & Willmes, 2005) and in different response setups (e.g., manual responses, pedal responses, saccadic eye movements; Schwarz & Keus, 2004; Schwarz & Müller, 2006), offline as well as online (Cipora et al., 2019; Roth, Jordan, et al., 2024). The association of number magnitude and space therefore seems to be highly robust and generalizable across many settings, even though many situational modulations have been described (Cipora, Patro, & Nuerk, 2018).

Importantly, the SNARC effect arises in several tasks, which – as we will outline in detail below – have major conceptual differences in the underlying semantic features of the numbers being processed. Two tasks that inquire about semantic numerical attributes of the digits/numbers themselves are by far the most frequently used to investigate the SNARC effect: (i) the magnitude classification (MC) task and (ii) the parity judgment (PJ) task (see Table 2 in Wood et al., 2008). In MC, participants judge whether numbers are smaller or larger than a reference number[1]. In PJ, participants judge whether numbers are odd or even. Although some studies have used other kinds of stimuli (e.g., dice patterns or number words in Nuerk et al., 2005, and multi-digit numbers in Tlauka, 2002; Weis et al., 2018), single-digit (i.e., Arabic or Hindu-Arabic) numbers are most used in this task. In both tasks, the instruction is to respond as

[1] In the current article, tasks in which presented numbers are to be compared with a fixed reference number (e.g., comparing whether a presented number between 1 and 9 excluding 5 is smaller or larger than 5) are referred to as magnitude *classification*. In contrast, tasks in which the reference number is not fixed but varies between trials are referred to as magnitude *comparison*.

quickly and as accurately as possible with a left- or right-hand key to numbers presented centrally on a computer screen. Typically, symbolic[2] numbers from 1 to 9 (excluding 5) are used as the stimulus set, with number 5 serving as the reference number in MC (Wood et al., 2008). In both tasks, the response-to-key assignment is flipped in the middle of the experiment, so that both left- and right-hand responses are given for each number.

In the following, we will first outline conceptual differences between MC and PJ (e.g., the relevance of number magnitude and number parity, SNARC and MARC compatibility, and the strength of the SNARC effect in MC and PJ) and a potential correlation between the MC- and PJ-SNARC. Most crucially, we will then discuss previously observed shape differences, and reasons for a categorical MC-SNARC. Additionally, we will introduce the potential influence of task order on the SNARC effect. At the end of this introduction, we will give an overview of all replication checks, hypotheses, and exploratory tests we aimed at testing in the current study.

Conceptual differences between MC and PJ

At first sight, apart from the instructions, the MC and PJ tasks seem to be similar: A semantic feature of numbers (i.e., magnitude or parity) is to be categorized. In the current study, we will use the most common bimanual computerized setup with symbolic numbers from 1 to 9 (excluding 5) described above. One might assume that the required cognitive processes, the responses given by participants, and the arising spatial mapping of number magnitude are similar in both tasks. However, as we shall see, the SNARC effects in MC and PJ (called *MC-SNARC* and *PJ-SNARC* in the remainder of this article) differ. In the following, we describe conceptual differences between MC and PJ to shed light on reasons for the different SNARC effects.

[2] Symbolic numbers consist of symbols, such as digits, whereas non-symbolic numbers do not consist of symbols but instead quantities, such as dot arrays.

Relevance of number magnitude and number parity

Most obviously, number magnitude is directly task-relevant to MC but not to PJ. This leads to an important difference: Tzelgov et al. (2015) distinguish between intentional and automatic processing, where the latter means processing without conscious monitoring. Automatic processing can be measured in tasks where the process in question is not part of the task requirements (Tzelgov, 1997). In contrast, intentional processing is supposed to reflect the task requirements.

For participants to show the SNARC effect, two representations have to be activated, namely the *magnitude* (e.g., two) or *ordinality* (e.g., the second) of a number and its directional association with space (Cipora, He, & Nuerk, 2020). Importantly, when judging parity, neither the processing of magnitude nor of its directional association with space is task-relevant and intentional. As a result, the PJ-SNARC effect is commonly regarded as an indicator of automatic number magnitude processing in humans. It also supports the notion that single-digit numbers function as primitives in Western cultures, such that their meanings can be effortlessly and holistically retrieved from memory without additional processing (Tzelgov et al., 2015). This is fundamentally different for MC, where the processing of magnitude is intentional, as it is task-relevant. Therefore, the MC-SNARC only shows that the directional association with space is automatic (as it is not needed for the response), but not that number-magnitude processing is automatic.

Importantly, the two tasks differ not only regarding task-relevance of number magnitude but also in the task-relevance of number parity. Parity is task-relevant in PJ while being task-irrelevant in MC. Apart from the SNARC effect, a phenomenon to be considered in the current study is the MARC effect (Linguistic Markedness of Response Codes; Nuerk et al., 2004), which is typically observed in PJ but not in MC (Replication Check 2; see ordered list of replication checks in the Introduction section

“The current study”). The MARC effect reflects faster responses to odd/even numbers on the left/right side, respectively. In a similar vein as for the SNARC effect, there are two prerequisites for the MARC effect: the processing of parity and its directional association with space. Crucially, the processing of number parity is less automatic than the processing of number magnitude (Roth et al., 2025a) and is more consistently found when using number words rather than Arabic digits (Nuerk et al., 2004; Roettger & Domahs, 2015). The processing of number parity seems not to be automatic, as typically no MC-MARC is found (Cipora, 2014; Deng et al., 2018). In contrast, the directional association of parity with space can be considered automatic, because it is not required in PJ, yet a PJ-MARC can typically be observed.

Further, response times (RT) increase when numerical magnitude increases, which is referred to as the Numerical Size Effect (Brysbaert, 1995). We expect the effect to arise in both tasks (Hypothesis 3a; see ordered list of hypotheses in the Introduction section “The current study”), although it has mainly been demonstrated for two-digit numbers (Brysbaert, 1995) and single-digit numbers will be used in the present study. Moreover, we expect it to be stronger in MC than in PJ (Hypothesis 3b) because numerical size (i.e., number magnitude) is only task-relevant in MC. Moreover, RT increase with increasing numerical distance between the stimulus and the reference number in MC, which is referred to as the Numerical Distance Effect (Moyer & Landauer, 1967). We expect the effect to arise in MC (Replication Check 4), but it cannot arise in PJ because there is no criterion that numbers are compared to.

SNARC and MARC compatibility

Importantly, the SNARC and MARC effects can also be considered to be *compatibility effects* (e.g., see cognitive-control account for the SNARC effect by Zhang et al., 2022). Crucially, MC and PJ differ with respect to compatibility. In a typical PJ task with two blocks, one block is MARC-compatible (i.e., when the instruction is to

respond to odd/even numbers with the left-/right-hand key, respectively) and the other block is MARC-incompatible. At the same time, within each block, half of the trials are SNARC-compatible (i.e., when the response to the parity of small/large numbers is assigned to the left/right, respectively) and half of the trials are SNARC-incompatible. On the contrary, in a typical MC task with two blocks, one block is SNARC-compatible (i.e., when participants are asked to respond to numbers that are smaller/larger than a reference number with the left-/right-hand key, respectively) and the other is SNARC-incompatible. SNARC-compatible and -incompatible trials alternating within blocks, as is the case in PJ, can elicit Gratton effects (Gratton et al., 1992). In line with this, Pfister et al. (2013) reported reduced SNARC effects after SNARC-incompatible compared to after SNARC-compatible trials in PJ. In a typical MC task with only two blocks, where SNARC-compatibility is grouped by block, such trial-to-trial effects cannot occur.

Apart from trial-to-trial compatibility effects, there can be compatibility effects depending on block order. Van Galen and Reitsma (2008) observed a stronger MC-SNARC in children who completed the SNARC-compatible followed by a SNARC-incompatible block than in children who were administered the reverse order, whereas Bulut et al. (2025) observed the opposite effect in one of three tested adult samples, but no effect in the two remaining adult samples. No effect of block order on the PJ-SNARC has been found (Bulut et al., 2025; Cipora, van Dijck, et al., 2019; Moorkens et al., 2025; Roth, Jordan, et al., 2024), where each block is either MARC-compatible or MARC-incompatible, while it consists of half SNARC-compatible and half SNARC-incompatible trials. For the MARC effect, influences of block order have been found in both directions. In a previous study, we found a stronger MARC effect in PJ with the MARC-incompatible-compatible order compared to the MARC-compatible-incompatible order (block-wise compatibility order differing between subjects, Cipora, van Dijck, et al., 2020). In another previous study, we found the reversed pattern

(block-wise compatibility order differing within subjects, Roth, Jordan, et al., 2024). The difference might be attributable to the design, and as block-wise compatibility order will differ between subjects in the current study like in the study by Cipora, van Dijck, et al. (2020), we expect the same pattern here. A stronger PJ-MARC in the MARC-incompatible-compatible order than in the MARC-compatible-incompatible order seems plausible: Participants need to familiarize themselves with PJ and overcome their natural odd-left and even-right association in the first block, while they are already familiar with PJ and can respond in line with their natural odd-left and even-right association in the second block. Participants therefore have two reasons to be slower in the first block, and the difference (i.e., the MARC effect) between blocks is therefore especially strong in this block order. We will exploratorily investigate compatibility-order effects in both tasks in the current study (Exploratory 2; see ordered list of exploratory tests in the Introduction section “The current study”).

Strength of the SNARC effect in MC and PJ

As outlined above, the processing of number magnitude is highly automatized and single-digit numbers can therefore be considered as primitives in Western cultures’ numerical cognition (Tzelgov et al., 2015). In contrast, the processing of number parity might be less automatized when only a semantic number feature other than parity is task-relevant (no MC-MARC found by Cipora, 2014, and Deng et al., 2018). Note that evidence has been found for both the MARC effect and the Odd effect (i.e., slower responses to odd than to even numbers, Hines, 1990) when only non-semantic features of numbers were judged (e.g., font color; Roth et al., 2025a), reflecting automatic parity processing to some extent, but the evidence was only weak. Also, number magnitude is more often relevant in daily life than number parity. Hence, the processing of magnitude seems to be more straightforward than the processing of parity. In line with this, average responses are typically faster in MC than in PJ (Kiesel

et al., 2007; Saeki & Saito, 2009; descriptively also observed by Fattorini et al., 2015; Fitousi et al., 2009; Gevers, Verguts, et al., 2006; Ito & Hatta, 2004, see also Wood et al., 2008, for a meta-analysis), which we expected to find in the current study as well (Replication Check 3).

The processing of magnitude being explicitly required in MC, but not in PJ, might elicit a stronger spatial mapping in MC than in PJ. In line with this assumption, the MC-SNARC has been found to be stronger than the PJ-SNARC (Bae et al., 2009; Cheung et al., 2015; Fitousi et al., 2009; van Dijck et al., 2009). On the other hand, judgments of number magnitude are automatic (Tzelgov et al., 2015) and therefore naturally faster than judgments of number parity. At the same time, the SNARC effect is typically stronger in slower responses in both MC and PJ, both within participants and on the sample level (see Table ST4 in Supplementary Materials in Cipora, Soltanlou, et al., 2019; see Table 3 in Didino et al., 2019; Gevers, Verguts, et al., 2006, Figure 6 there). In contrast to the reasoning above, this would lead to the opposite prediction of the PJ-SNARC being stronger than the MC-SNARC, which has been observed by Georges et al. (2017), Gevers, Verguts, et al. (2006), and Ito and Hatta (2004). No difference between the MC-SNARC and the PJ-SNARC was found by Didino et al. (2019) in an independent-samples *t*-test. Taken together, (i) easier and possibly deeper processing of magnitude in MC than in PJ should lead to a greater SNA in MC, and (ii) longer response times in PJ than in MC should lead to a greater SNA in PJ. Both opposing processes seem to be valid and there is no clear picture in the literature. It remains unclear whether the SNARC effect differs in size between tasks, and we therefore look at this in an exploratory analysis (Exploratory 6).

Further differences between MC and PJ

Several more differences may exist between MC and PJ. First, the MC-SNARC seems to more strongly involve visuospatial working memory, the PJ-SNARC seems

to rely more on verbal working memory (Deng et al., 2017; Herrera et al., 2008; van Dijck et al., 2009). Second, the MC-SNARC and the PJ-SNARC might arise at different processing stages (Basso Moro et al., 2018; Xiang et al., 2022). Third, cognitive mechanisms underlying the MC-SNARC and the PJ-SNARC might differ. Namely, Prpic et al. (2016) claim that ordinality drives the SNARC effect in *direct* tasks (e.g., MC, where magnitude is response-relevant), whereas cardinality underlies in *indirect* tasks (e.g., PJ, where magnitude is response-irrelevant). Note that Casasanto and Pitt (2019) claim that only ordinality is crucial for both direct and indirect tasks, and that Koch et al. (2023) show that order- and magnitude-related mechanisms are not mutually exclusive. Looking into these differences between the MC-SNARC and the PJ-SNARC is beyond the scope of the current study; however, the current study provides an improved understanding of the two tasks and thereby lays the groundwork for further investigation.

In summary, several conceptual differences exist between MC and PJ, concerning the task-relevance of number magnitude and number parity, the compatibility of the response-to-key assignment with the SNARC and MARC effects, arising numerical-cognition effects, and underlying cognitive mechanisms. However, both tasks elicit a SNARC effect, and the presence and strength of the MC- and PJ-SNARC in the same participants might be related to one another, as discussed in the next section.

Correlation between the MC- and PJ-SNARC

After having described the similarities and differences of MC and PJ, the question arises whether the MC-SNARC and the PJ-SNARC are correlated. However, several factors might lead to a null correlation. First, there seem to be high fluctuations in the SNARC effect over time (Roth, Jordan, et al., 2024) that limit the maximum correlation that can be detected. Second, the test-retest reliability of the SNARC effect

has been found to be poor for both MC and PJ (correlations .22 < r < .41; Cipora & Göbel, 2013; Georges et al., 2013; Hedge et al., 2018; Viarouge et al., 2014). The lower the test-retest reliabilities of the MC-SNARC and PJ-SNARC, the lower is also the maximally observable correlation between the two effects. Third, the split-half reliability of the SNARC effect has been found to be poor for PJ at least in some studies (correlations .43 < r < .96; for an overview, see Table 1 in Cipora, van Dijck, et al., 2019). To conclude, both properties of the SNA construct and its operationalization in experimental tasks influence whether a correlation between the MC-SNARC and the PJ-SNARC is found. Possible reasons for a null finding could be low intraindividual stability, low reliability, or low internal consistency, whereas we would interpret a high correlation between the MC-SNARC and the PJ-SNARC as both MC and PJ reliably measuring the same underlying theoretical construct.

Several previous studies did not find any correlation (correlations with 95% confidence intervals and *p*-values: r = -.02 [-0.19, 0.15] and p = .822 for Germans; r = -.08 [-0.26, 0.10] and p = .386 for Turks; r = .10 [-0.13, 0.32] and p = .402 for Iranians, in Bulut et al., 2025; r = .09 [-0.18, 0.35] and p = .513 in Cipora, 2014; r = 0.06 [-0.30, 0.40] and p = .744 in Didino et al., 2019; r = .18 [-0.07, 0.42] and p = .18 in Fattorini et al., 2015; r = 0.20 [-0.01, 0.39] and p = .07 in Georges et al., 2017). To our knowledge, a statistically significant correlation has only been reported by Cheung et al. (2015; r = 0.25) and by Cipora (2014; r = .50, but only in a unimanual setup). Note that an existing weak correlation between the MC-SNARC and the PJ-SNARC despite limiting factors such as low intraindividual stability, low reliability, or low internal consistency would only be detectable in large samples. To be able to detect a potential correlation between the MC-SNARC and the PJ-SNARC, we administered both tasks to a large sample in a within-subjects design. This enabled us to test the correlation between the

MC-SNARC and the PJ-SNARC in an exploratory analysis with high statistical power (Exploratory 7).

Different shapes of the SNARC effect

After having outlined the differences between MC and PJ and after having discussed the potential correlation between the SNARC effect in these two tasks, it is important to note that the shape of the MC-SNARC seems to be categorical, while the shape of the PJ-SNARC seems to be continuous (Wood et al., 2008). However, the MC-SNARC is often modelled as a continuous phenomenon, just as the PJ-SNARC, as described in the following. The main aim of the current study is to thoroughly investigate the SNARC effect in these two most widely used tasks and to find out how to best statistically model the SNARC effect.

The SNARC effect is usually calculated by subtracting the mean RT with the left hand from those with the right hand for each number and regressing these differences (dRT) on magnitude as a continuous predictor in both MC and PJ. A negative regression slope reflects the increasing right-hand advantage for larger numbers and therefore the SNARC effect. To investigate whether the effect is present on group level, regression slopes (one per participant) are then tested against zero in a one-sample *t*-test (repeated-measures regression, adapted by Fias et al., 1996, based on Lorch and Myers, 1990). Importantly, this analysis method is only suitable for a continuous SNARC effect, reflecting a constant increase in right-hand advantage (reflected by a constant decrease in dRT) per increase of magnitude. Hence, when participants judge whether numbers in the typically used stimulus set from 1 to 9 (excluding 5) are odd or even, the spatial mapping of extreme magnitudes such as 1 and 9 is stronger than for magnitudes closer to the mid of the stimulus set such as 4 and 6. In other words, the association with the left side is stronger for the very small number 1 than for the slightly small number 4. While the PJ-SNARC is linear, the MC-SNARC is typically categorical

(e.g., Gevers, Verguts, et al., 2006), especially in adults (van Galen & Reitsma, 2008): In a typical MC task, when averaging across all participants, responses are *equally* faster with the left hand to numbers from 1 to 4 and equally faster with the right hand to numbers from 6 to 9. Therefore, a stepwise model reflects the MC-SNARC better than a continuous model (as reflected by a better model fit in terms of a higher proportion of explained variance). The use of a categorical instead of a linear function for quantifying the MC-SNARC would increase the model fit at the participant level and thereby likely also the precision of the effect size estimate at the sample level.

Nevertheless, a linear predictor in the regression of dRT on number magnitude remained a frequently used analysis of the MC-SNARC (Bachot et al., 2005; Bae et al., 2009; Bull et al., 2005; Cheung et al., 2015; Deng et al., 2017; Han et al., 2017; Herrera et al., 2008; Hoffmann et al., 2013; E. M. Hubbard et al., 2009; Ito & Hatta, 2004; Lohmann et al., 2018; Mourad & Leth-Steensen, 2017; Nathan et al., 2009; Pinto et al., 2021; Schiller et al., 2016; Shaki & Gevers, 2011; van Dijck & Doricchi, 2019; van Dijck et al., 2009; van Dijck et al., 2012; van Galen & Reitsma, 2008; Weis et al., 2018). As the correlation between the linear predictor (i.e., 1, 2, 3, 4, 6, 7, 8, 9) and a categorical predictor (especially the most commonly assumed one, i.e., -0.5, -0.5, -0.5, -0.5, 0.5, 0.5, 0.5, 0.5) is extremely high (r = .913), the model with a linear magnitude predictor fits relatively well both to the continuously and categorically distributed dRT in MC and PJ (see top panel of Figure 1 in Bae et al., 2009, or Figure 1 in Nathan et al., 2009). In some studies, a two-way ANOVA including magnitude (small vs. large) and response side (left vs. right) as within-subjects factors has been used to quantify the MC-SNARC (Fattorini et al., 2015; Gevers, Verguts, et al., 2006; Herrera et al., 2008; Hoffmann et al., 2013; Nathan et al., 2009). However, compared to that approach, the repeated-measures regression approach has several advantages (Fias et al., 1996): First, the presence of the SNARC effect is judged by a

main effect instead of an interaction effect, which allows a quantification of the size of the effect in milliseconds by the slope. Second, the presence or absence of a SNARC effect can be assessed for each participant individually. A repeated-measures regression with a categorical predictor for the MC-SNARC has only been used in few studies (Bulut et al., 2025; Cipora, 2014; Didino et al., 2019; Fitousi et al., 2009; Georges et al., 2017; Gevers, Verguts, et al., 2006; Hohol et al., 2020; Nathan et al., 2009; Nuerk, Wood, & Willmes, 2005; Weis et al., 2018; Zorzi et al., 2012). For an overview of all mentioned studies including MC, see Table A1 in Appendix A. The categorical predictor was defined using 5 as a boundary (so that 1 to 4 are considered as “small” and 6 to 9 as “large”) in all the named studies, and, if not specified otherwise, the same categorical predictor is meant in the current article (for alternative considerations, see below in Exploratory 4).

Unfortunately, the suitability of the linear and categorical predictors for dRT in MC with the stimulus set from 1 to 9 (excluding 5) was assessed by direct comparison in only a few studies. Fitousi et al. (2009) and Nathan et al. (2009) computed two separate regression models, one of which with a categorical and the other with a linear predictor, and in both studies the fit was higher with a categorical (R^2 = .904 in Fitousi et al., 2009; R^2 = .988 in Nathan et al., 2009) than with a linear predictor (R^2 = .775 in Fitousi et al., 2009; R^2 = .891 in Nathan et al., 2009). Similarly, Didino et al. (2019), Gevers, Verguts, et al. (2006) and Nuerk, Bauer, et al. (2005) ran regression analyses including both predictors, and only the categorical predictor turned out to be significant in all three studies. In a study with two-digit numbers where participants performed PJ and MC for the unit digit, Weis et al. (2018) also included both linear and categorical predictors for both unit and decade magnitude concurrently into one regression model. They found only a categorical predictor for units and decades to be significant in MC and only the linear predictors for units and decades to be significant in PJ, providing

further evidence for a categorical MC-SNARC and a continuous PJ-SNARC. Importantly, because a linear model fits well even for the categorical MC-SNARC (e.g., see Fitousi et al., 2009; and Nathan et al., 2009), evidence for a better fit of a categorical model in MC can only be achieved with sufficient power by using a large sample and a sufficient number of repetitions per experimental cell (resulting from the combination of each stimulus with each response hand per task). As outlined above, although a linear model fit to the MC-SNARC might be high in some studies, a categorical model seems to be more adequate. We expected to find a better fit of a categorical model in MC (Hypothesis 1) and of a linear model in PJ (Hypothesis 2).

Importantly, the boundary between “small” and “large” numbers is not necessarily 5 for everyone and might instead vary between individuals. A split of the full number interval into two halves (i.e., from 1 to 4 and from 6 to 9) seems plausible in MC, where the boundary of 5 is explicitly defined in the task instructions. However, especially in PJ, but potentially even in MC, some individuals might classify numbers into the categories “small” and “large” with a different boundary.[3] Crucially, an overall continuous SNARC effect could result from continuous patterns in most individuals, but also from averaging across categorical patterns differing between individuals. In this study, we therefore additionally determined the most likely categorical boundary for each participant separately. Subsequently, we investigated the shape of the SNARC effect (Hypotheses 1 and 2) once more, this time comparing the best fitting categorical model for each individual with the continuous model (Exploratory 4).

Explanations for the categorical MC-SNARC shape

The literature provides several explanations for the different shapes of the SNARC effect, depending on the task. First, numbers are typically roughly classified

[3] We wish to thank the stage-1 PCI-RR reviewer Peter Wühr for bringing this conceptual alternative to our awareness.

into small and large numbers (Banks et al., 1976; Tzelgov et al., 1992), which is sufficiently precise to perform MC and might lead to the categorical MC-SNARC (Fitousi et al., 2009; Gevers, Verguts, et al., 2006). In contrast, participants are not instructed to process number magnitude at all in PJ, and thus number magnitude processing is not intentional (i.e., slow conditional route according to the dual-route model by Gevers, Ratinckx, et al., 2006) but rather automatic (i.e., fast unconditional route)[4]. Automatic number magnitude processing seems to be more exact and is continuously mapped onto space in PJ. This explanation is in line with the polarity-correspondence account of the SNARC effect by Proctor and Cho (2006), as well as with the application of the markedness principle to number magnitude (Nuerk & Schroeder, 2024; Schroeder et al., 2017). According to these two theories, the SNARC effect arises because both *large* and *right* are associated with the positive or unmarked polarity and both *small* and *left* with the negative or marked polarity. Similarly, this explanation is compatible with the verbal-spatial account of the SNARC effect proposed by Gevers, Verguts, et al. (2006; see also Gevers et al., 2010), stating that verbal categories such as *small* vs. *large* and *left* vs. *right* are responsible for the SNARC effect. These accounts argue for an intermediate classification into small or large numbers (Santens & Gevers, 2008), and the polarities, markedness, or verbal labels are categorical rather than continuous (Bae et al., 2009), which explains the categorical MC-SNARC. Note that it is possible that the PJ-SNARC is linear in the beginning of the task and becomes categorical over the course of the task, so that the continuous shape shifts to a stepwise one. That is, participants might start classifying

[4] As outlined by the stage-1 PCI-RR reviewer Peter Wühr, predictions can be derived from the dual-route model. Importantly, both the automatic and the intentional route are activated in MC, whereas only the automatic route is activated in PJ. Thus, the SNARC effect should be stronger in MC than in PJ because it results from both routes instead of only one route. Moreover, a positive correlation of the MC- and PJ-SNARC can be assumed based on the dual-route model, since both effects are (at least partly) caused by the automatic route.

the stimuli into the two categories “small” and “large” in PJ as soon as they become familiar with the stimulus set because they might notice that the stimulus set consists of two single-digit number sequences (i.e., 1 to 4 and 6 to 9) separated by the missing number 5. We investigated this possibility in an exploratory analysis (Exploratory 5).

Second, the Numerical Distance Effect might play a role for the MC-SNARC (Gevers, Verguts, et al., 2006). For instance, if the reference number is 5 in MC with the stimulus set from 1 to 9, numbers 4 and 6 need to be processed more intensely than numbers 1 and 9 to discriminate them from number 5 (Wood et al., 2008). Hence, responses are slowest for numbers 4 and 6 and fastest for numbers 1 and 9 with this stimulus set. In combination with the finding that the SNARC effect becomes stronger with increasing RT, the absolute values of dRT for number magnitudes that are close to the reference are larger than dRT predictions by the linear SNARC regression slope, resulting in a categorical shape (Didino et al., 2019; Georges et al., 2017; Gevers, Verguts, et al., 2006). Importantly, the Numerical Distance Effect demonstrates automatic processing of number magnitude (i.e., fast unconditional route), because it reflects performance differences in the discrimination between numbers and arises although the task instructions do not favor or disfavor the performance for specific stimuli. It thus does not build on the gross classification into smaller and larger numbers described in the previous paragraph (i.e., slow conditional route), but rather on the exact number magnitudes.

In summary, the rationale for a categorical instead of linear MC-SNARC is twofold: First, the intentional classification into small and large numbers is categorical in MC, and second, the interaction between the Numerical Distance Effect and the positive correlation between the SNARC effect and overall RT contributes to a step-wise shape. Because statistical models should correspond to scientific models as

closely as possible (Westermann & Hager, 2017), the MC-SNARC should therefore be tested with a categorical predictor.

Influence of task order on the SNARC effect

The influence of task order (first MC and second PJ, or reversed) on the MC-SNARC and on the PJ-SNARC has been investigated in only a few studies. Didino et al. (2019) did not find any task-order effects on the SNARC effect. Fattorini et al. (2015) did not observe any task-order effect on the PJ-SNARC but found the MC-SNARC to be weaker after PJ than when MC was the first task. In most other studies that included the two tasks, effects of task order have either not been reported (Cheung et al., 2015; Gevers, Verguts, et al., 2006; Nuerk, Bauer, et al., 2005; Weis et al., 2018), or could not be calculated because task order was not counterbalanced (Bae et al., 2009; Cipora, 2014; Fitousi et al., 2009; Georges et al., 2017; Zorzi et al., 2012), or because different samples completed MC and PJ (Ito & Hatta, 2004; van Dijck et al., 2009). To our knowledge, only Bulut et al. (2025; see Supplementary Materials) have tested the influence of task order, and they did not find an influence on the SNARC effect in any of the two tasks in any of three samples (130 German, 112 Turkish, and 75 Iranian participants). Two opposite theoretical predictions can be made. On the one hand, the SNARC effect might be stronger in each task if it is the second, because the processing of number magnitude and its spatial mapping should be stronger if they have already been activated in a previous task. On the other hand, the SNARC effect might be weaker in each task if it is the second, because RT typically decrease with practice and faster RT are typically associated with a weaker SNARC effect (note that both decreasing RT and a decreasing SNARC effect over time in PJ have been found by Roth, Jordan, et al., 2024). If both mechanisms were true, they might cancel out each other and make the influence of task order invisible. Hence, we could not make any

directional prediction and will investigate the potential influence of task order in an exploratory analysis (Exploratory 1).

The current study

In this large-scale online study, we investigated whether the MC-SNARC is truly categorical (i.e., better described by a categorical number magnitude predictor) and the PJ-SNARC continuous (i.e., better described by a continuous number magnitude predictor). Evidence for this systematic difference would suggest that we should not call it *the* SNARC effect, but instead acknowledge that different SNARC effects exist, which are elicited depending on the task. It is crucial to shed light on this issue, because wrong measurements and interpretations of the SNARC effect can lead to misconceptions of SNAs. Another goal of the present study was to investigate the relationship between the SNARC effect(s) observed in the two tasks. In the following, all replication checks, hypotheses, and exploratory tests are listed (related ones follow one another).

First, we tested the following expectations as replication checks:

1. a SNARC effect in both MC and PJ with the standard analysis of a continuous linear regression (this positive control will be used as a basis for all further analyses, i.e., finding a SNARC effect with the standard analysis in both tasks is a prerequisite for testing the three hypotheses in this study);
2. a MARC effect in PJ, but not in MC, because the activation of parity seems not to be automatic if only a semantic number feature other than parity is task-relevant;
3. shorter RT in MC than in PJ, because processing magnitude is more straightforward and automatized than processing parity;
4. a Numerical Distance Effect in MC, which is typically found.

To summarize our three hypotheses derived above, we expected:

1. a categorical MC-SNARC, i.e., a better fit of a categorical than continuous MC-SNARC model
2. a continuous PJ-SNARC, i.e., a better fit of a continuous than categorical PJ-SNARC model;
3. (a) a Numerical Size Effect in both tasks, (b) which is stronger in MC than in PJ, because processing magnitude is task-relevant in MC but task-irrelevant in PJ.

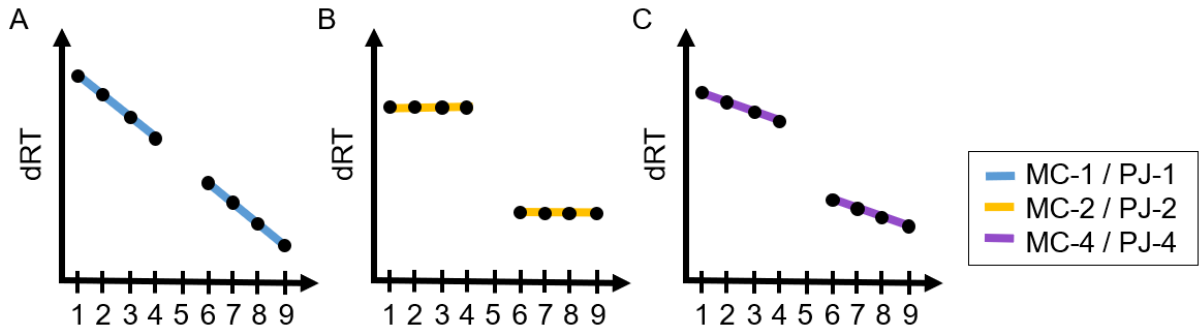
Moreover, we explored whether the following observations can be made:

1. task-order effects (Conditions 1 and 2 versus Conditions 3 and 4) on both (a) the MC-SNARC and (b) the PJ-SNARC;
2. compatibility-order effects on (a) the MC-SNARC (SNARC slopes in Conditions 1 and 3 versus Conditions 2 and 4) and on (b) the PJ-MARC (MARC slopes in Conditions 1 and 3 versus Conditions 2 and 4);
3. a good model fit when including both continuous and categorical magnitude predictors for (a) the MC-SNARC or for (b) the PJ-SNARC, indicating a mixed shape of the SNARC effect (see Panel C in Figure 1);
4. number 5 as the most likely boundary between “small” and “large” numbers for most individuals both in MC and PJ; and, relatedly, a categorical MC-SNARC (Hypothesis 1) and a continuous PJ-SNARC (Hypothesis 2), even when comparing the favored categorical model for each participant with the continuous model;
5. a shape difference of (a) the MC-SNARC or for (b) the PJ-SNARC between earlier and later phases within each task;
6. a stronger MC-SNARC than PJ-SNARC when the continuous magnitude predictor is used;

7. a positive correlation between categorical MC-SNARC slopes and continuous PJ-SNARC slopes.

Figure 1

Different shapes of the SNARC effect



Note. In Panel A, the SNARC effect is reflected by a linear regression line with a negative slope, that is, dRT are best predicted by continuous magnitude (models MC-1 and PJ-1). In Panel B, the SNARC effect is reflected by a step-wise function, that is, dRT are best predicted by categorical magnitude (models MC-2 and PJ-2), where all numbers smaller than 5 have the same dRT and all numbers larger than 5 have the same dRT. Panel C shows an intermediate shape of the SNARC effect, where both continuous and categorical magnitude predict dRT (models MC-4 and PJ-4). The typically observed MC-SNARC appears as shown in Panel B (Hypothesis 1), and the typically observed PJ-SNARC appears as shown in Panel A (Hypothesis 2). The figure was created by L. Roth using PowerPoint (Microsoft Corporation, 2024).

We collected data for MC and PJ with the numbers from 1 to 9 (excluding 5) in a within-subjects design using the typical bimanual response setup. Participants were assigned to one of four conditions differing in block order, and 30 repetitions provided reliable estimates per experimental cell (number magnitude * response side * task; see Cipora & Wood, 2017). Conducting this study online offered the possibility to test much larger samples than in most previous studies and thus reach high statistical power (Reips, 2000, 2002). The SNARC effect and its various forms have been successfully

replicated in online settings (Bulut et al., 2025; Cipora, Soltanlou, et al., 2019; Gökaydin et al., 2018; Koch et al., 2023; Roth et al., 2025a; Roth et al., 2025b; Roth et al., 2025c). The measurement in the online setup showed reliability and a similar magnitude compared to the SNARC effect that is typically observed in lab studies. Further, it seems to be valid regarding correlations with mean RT and standard deviations of RT. We calculated Bayes Factors (BF_{10}) to quantify evidence both for differences between MC and PJ as well as for the relationship between the SNARC effects in the two tasks, and lack of such differences or such a relationship. This way, we could shed more light on different SNARC effects specifically in terms of their shape in the two most popular and widely used tasks.

Method

The ethics committee of the University of Tübingen's Department of Psychology approved this study. After peer review and prior to data collection, the Stage-1 Registered Report received in-principle acceptance by the Peer Community In on March 3rd, 2025.

Sample size considerations

The "Sequential Bayes Factor with maximal n" (SBF+maxN) approach described by Schönbrodt and Wagenmakers (2018)[5] was applied to make our data collection efficient. This means that we ran the data analysis with a total of 500 participants in the first round and recruited further participants sequentially in steps of 50 until the stopping criterion or maximal sample size was reached. Our stopping

[5] Note that, apart from a maximum sample size, a minimum sample size needs to be defined as well, which is why it might be more reasonable to term the approach "sequential Bayes Factor with a minimum and maximum N" as done by Witt (2019).

criterion was moderate evidence regarding all three hypotheses, so that the data either provides evidence in favor of (BF_{10} > 3) or against (BF_{10} < 1/3) each of them.

For the SBF+maxN approach, we defined a maximal (and a minimal) sample size. Thus, we determined the sample size that was necessary to detect evidence for a true underlying effect or for true absence of an effect with a high probability (similar to power analyses in the frequentist framework). This was done for each of the three hypotheses and the largest required sample size was chosen as maximal sample size for the SBF+maxN approach. Our main aim of the current study was to determine the shape of the SNARC effect in each of the two most common tasks. For this, we compared the fit of a continuous and a categorical statistical model in MC and PJ separately (to test Hypotheses 1 and 2). We therefore chose the effect size of interest (ESOI) in a standardized unit, namely Cohen's d = 0.2 (although this is not ideal for power simulations, see Correll et al., 2020). The sample size considerations were based on this ESOI for all three hypotheses, because smaller effect sizes are not practically meaningful. Specifically, d = 0.2 reflects a small effect and corresponds to around 1% of explained variance (calculated using the conversion formula assuming equal-sized groups in Table 2 in Ruscio, 2008). Regarding the detection of a SNARC effect while assuming similar standard deviations as reported in the literature, d = 0.2 corresponds to slopes of -4 ms in the continuous MC-SNARC (with SD = 20), -10 ms in the categorical MC-SNARC (with SD = 50), and -2 ms in the continuous PJ-SNARC (with SD = 10) in their measured unit (i.e., increase of right-hand advantage per continuous magnitude or categorically for large compared to small numbers in milliseconds). These SNARC slopes are of a typically observed or even smaller size.

Analogously to statistical power simulations in the frequentist framework, we randomly drew 5000 samples from a distribution around the ESOI (d = 0.2) and simulated the probability to obtain at least moderate evidence (i.e., BF_{10} > 3) for that

effect size by looking at the proportion of Bayesian tests revealing at least moderate evidence for the alternative hypothesis (for a similar approach, see Kelter, 2021; Roth et al., 2025b). Similarly, we randomly drew 5000 samples from a distribution with the respective SD around a truly absent effect ($d = 0$) and simulated the probability to obtain at least moderate evidence for the null hypothesis (see Kelter, 2021). We thereby determined the sample size that was required for a probability of .90 to obtain moderate evidence for and against the three hypotheses with two-sided Bayesian paired or one-sample *t*-tests. The required sample size was set as largest for finding at least moderate Bayesian evidence for a true underlying effect of $d = 0.2$ with a probability of .90 in a two-sided Bayesian independent-samples *t*-test ($n = 2 * 850 = 1700$). The required sample sizes for finding evidence showing true absence of an effect in an independent-samples *t*-test ($n = 2 * 340 = 680$), for evidence for a true underlying effect in a one-sample or paired *t*-test ($n = 440$), or for evidence showing true absence of an effect in a one-sample or paired *t*-test ($n = 160$) were much smaller. We therefore chose $n = 1700$ as a maximal sample size for the SBF+maxN approach. The exact calculations and results for all tests can be found here: https://osf.io/4wpv6/.

Participants

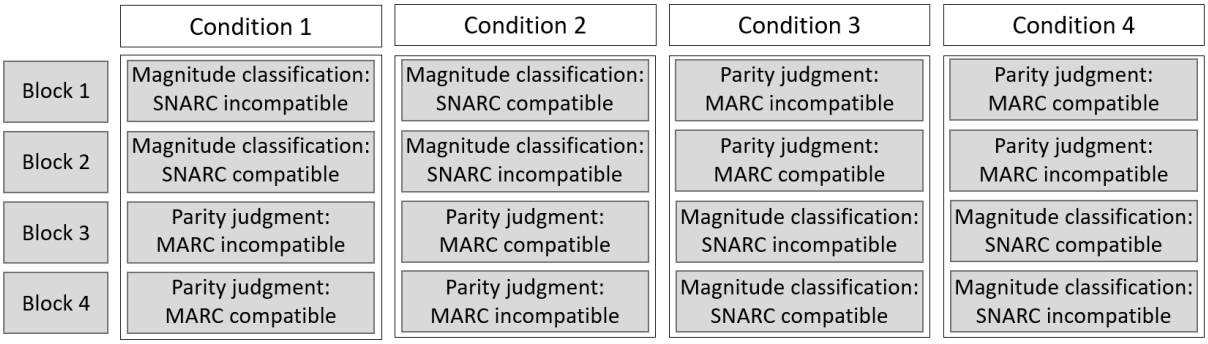
We sequentially recruited adults aged between 18 and 40 years via the recruiting platform Prolific (note that participants' age is self-reported during signup, which can be an issue in recruitment for online experiments, see Reips, 2021). As the study was conducted in English, participation was only possible for native English speakers (as per Prolific's screening based on self-reports). Complete participation was compensated with £5 (Prolific users receive their payment in Great British Pounds), and incomplete participation was compensated partially.

Design and experimental task

The present study followed a 2 (task: MC vs. PJ) * 2 (compatibility: incompatible vs. compatible) within-subjects design, resulting in four experimental blocks per participant. Participants were randomly assigned to one of four block orders that resulted from the constraint that both compatibility versions of the same task should follow directly after each other to avoid mixing up instructions (cf. Figure 2). In Conditions 1 and 2, participants completed MC in the first and PJ in the second half of the experiment, while the task order was reversed in Conditions 3 and 4. Within each task, participants were assigned to the SNARC-/MARC-incompatible block first and to the SNARC-/MARC-compatible block second in Conditions 1 and 3, while the compatibility order was reversed in Conditions 2 and 4. Given the number of assessed trials (see below), each of the two tasks took approximately 15 minutes, resulting in a median time needed for full participation (including both tasks and some demographic and other questions like the seriousness check) of 33 minutes and 58 seconds.

Figure 2

Within-subjects manipulations and block orders counterbalanced across participants



Note. We randomly assigned participants to one of the four conditions resulting from the combination of task order and compatibility order in the 2 (task: MC vs. PJ) * 2 (compatibility: incompatible vs. compatible) within-subjects design. The figure was created by L. Roth using PowerPoint (Microsoft Corporation, 2024).

A binary response-key setup was employed, requiring participants to respond as quickly and accurately as possible using a left or right key (defaults: D or K – can be adjusted individually by participants due to large technical variance on the Internet; Reips, 2000, 2021) depending on whether the number presented on the screen was smaller or larger than 5 (MC) or whether it was odd or even (PJ). In each of the experimental conditions resulting from the two within-subjects factors *task* and *compatibility*, number magnitude (1, 2, 3, 4, 6, 7, 8 vs. 9) was manipulated. Thirty repetitions per experimental cell lead to 240 SNARC-incompatible and 240 SNARC-compatible trials in MC, as well as 240 MARC-incompatible and 240 MARC-compatible trials in PJ per participant. Participants had to take a break of a minimum of 30 seconds between blocks. The order of stimuli within each block was randomized, but each stimulus was presented for the 1st throughout 15th time before any were presented for the 16th throughout 30th time. Each block was divided into two subblocks indistinguishable to the participant, in which each stimulus was presented 15 times (note that this was necessary to investigate Exploratory 5). Each trial started with a square (extended ASCII 254, size 72px), serving as the eye fixation point (300 ms), presented in the center of the screen. Then the number (Open Sans font, size 72px) replaced the square and remained on the screen until a response was given. A blank gray screen (500 ms) concluded the trial. Stimuli as well as fixation squares were presented in black color (0, 0, 0 in RGB notation), while the background remained gray (150, 150, 150 in RGB notation) throughout the experiment. A practice session with 16 trials preceded each block, in which each number was presented twice. Accuracy feedback appeared during practice sessions only.

Procedure

At the beginning of the experiment, a seriousness check (e.g., Reips, 2009) was applied, i.e., participants were asked whether they intended to participate seriously.

Participants were asked to take part only if they gave their informed consent, if they used a computer (participation from mobile devices was not possible because a keyboard was required), and if they were between 18 and 40 years old. Then, participants were asked to provide basic demographic data, namely age, gender, first native language, and handedness. They were also asked about their finger counting habits, i.e., which hand they usually start counting with (*right*, *left*, *do not know or no preferred hand*) and how stable their preference is (*always*, *mostly*, *slightly more often than with the other hand*, *do not know or no preferred hand*). In each question, participants had the opportunity to click on "I prefer not to answer." Next, participants could choose response keys for the experimental task that are located on the same height and about one hand width apart from each other on their keyboard. The default keys D and K meet these requirements on typical keyboards like QWERTZ, QWERTY, and AZERTY and participants were asked to reject that default and change keys if they used a different keyboard layout. . Then, instructions were displayed, and the first block of the experimental task started with its practice trials.

After completion of both experimental tasks, data quality was assessed by asking participants how they would describe their environment during participation (*silent*, *very quiet*, *fairly quiet*, *fairly noisy*, *very noisy,* or *extremely noisy),* whether there were any major distractions during participation (*none*, *one,* or *multiple*), and whether there were any difficulties during participation (*yes* or *no*, text field for comments). Moreover, we asked participants whether they had used their left and right index fingers throughout the experiment, as asked for in the task instructions (*yes*, *partly*, or *no*). Participants were provided with a completion code to be inserted in Prolific and with our research team's contact information.

The experiment was set up with WEXTOR (https://wextor.eu; Reips & Neuhaus, 2002) in its HTML and JavaScript framework and adapted (see demo version at

https://exp.wextor.eu/esnarc/task/?demo). Our previous experiments have demonstrated that this software is suitable for detecting the SNARC effect in an online setup (Roth et al., 2025a; Roth et al., 2025b). To prevent search engine bots (e.g., Googlebot) from submitting data to our experiment, WEXTOR equips the experiment materials with a standardized “noindex, nofollow” meta tag, which prompts search engine bots not to index the experiment pages and also not to visit subsequent pages (see Reips, 2007, p. 379). Further, we restricted participation to devices with a screen width of over 600 pixels. Additionally, to exclude multiple submissions from the same devices, we performed checks based on User-Agents and IP addresses during data evaluation.

Data preprocessing

All data preprocessing steps and all statistical analyses were performed in the computing software R (R Core Team, 2022). As concerns data preprocessing, we stayed consistent with our previous studies and applied similar inclusion criteria (Roth et al., 2025a; Roth et al., 2025b). Only datasets of participants who completed both tasks, who indicated to be between 18 and 40 years old, and who stated their intention to seriously participate were analyzed. Datasets were not included for analyses if participants described their environment as very/extremely noisy, if they reported multiple major distractions, or if participants did not use their left/right index finger for the left/right response key, respectively. Practice trials and incorrectly answered trials were not analyzed. Only trials with RT from 200 to 1500 ms were included in the analysis. Further outliers were removed in an iterative trimming procedure for each participant and task separately, such that only RT that are a maximum 3 SDs above or below the individual mean RT of all remaining trials after these exclusions were considered. Finally, only datasets of participants with at least 75% valid remaining trials

per task and without any empty experimental cell (number magnitude * response side * task) were considered.

Data analysis

We calculated Bayes Factors associated with the corresponding Bayesian *t*-test to obtain evidence for both null and alternative hypotheses (using a default *r*-scale of 0.707 as uninformed prior using Cauchy distribution). A resulting BF_{10} greater than 3 or 10 was treated as moderate or strong evidence for the alternative hypothesis compared to the null hypothesis, respectively, while a resulting BF_{10} smaller than 1/3 or 1/10 was treated as moderate or strong evidence for the null hypothesis compared to the alternative hypothesis, respectively (Dienes, 2021). When using a default *r*-scale of 0.707 Cauchy prior distribution, classifying a BF_{10} larger than 3 as evidence against the null hypothesis is more conservative than rejecting a null hypothesis in the frequentist framework with the typical significance level of $\alpha = .05$ (Wetzels et al., 2011). As explained above, we applied the SBF+maxN approach for sequential data analysis with optional stopping in case of at least moderate evidence for or against each of the three hypotheses.

RT were measured as the time elapsing from the onset of the number presentation on the screen until a response key is pressed (within the limitations that apply in Internet-based research with consumer-grade equipment, see e.g., Garaizar & Reips, 2019). As the dependent variable, we calculated the dRT, which resulted from subtracting the average RT of the left hand from the average RT of the right hand for each number separately per participant and for each task separately.

Several regression models were fit for each participant separately. In these regression models, number magnitude was included as a predictor for dRT to determine the shape of the SNARC effect in each task separately. First, magnitude was included as a continuous predictor, which was equal to the actual stimulus

displayed (e.g., 3 for number 3, and 8 for number 8). The resulting regression slopes for *continuous magnitude* represent the advantage of right-hand responses compared to left-hand responses in ms per increase by one in continuous magnitude (i.e., traditional repeated-measures regression in the SNARC effect analysis, as first proposed by Fias et al., 1996). Second, magnitude was contrast-coded as a categorical predictor, using -0.5 for numbers from 1 to 4 and +0.5 for numbers from 6 to 9 (but see Exploratory 4 for a different approach). The resulting regression slopes for *categorical magnitude* represent the advantage of right-hand responses compared to left-hand responses in ms in large compared to small magnitude. Third, for the investigation of the MARC effect, contrast-coded number parity was included as a predictor of dRT, with -0.5 for odd and +0.5 for even numbers (as in Cipora, Soltanlou, et al., 2019). The regression slopes for *parity* represent the advantage of right-hand responses compared to left-hand responses in ms in even compared to odd numbers. An overview of all three predictors (i.e., continuous magnitude, categorical magnitude, and parity), along with their exact coding, can be found in Table 2. For each of the predictors, a more negative coefficient estimate β points towards a stronger SNARC/MARC effect.

Confirmatory data analysis

An overview of all replication checks and hypotheses, corresponding statistical tests, interpretations of possible outcomes, and corresponding theoretical conclusions is given in the Study Design Table (Table 1). The Study Design Table is based on a template from the Peer Community In (https://peercommunityin.org) and was filled in prior to data collection – only the rightmost column with truly observed outcomes was added post data collection.

Table 1

Study Design Table based on a template from the Peer Community In

Question	Hypothesis	Sampling plan[1]	Analysis Plan	Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes	Observed outcome
Can a continuous SNARC effect be replicated in magnitude classification (MC) and in parity judgment (PJ)?	**Replication Check 1:** A significant SNARC effect is observable in both MC and PJ when determined with the standard analysis of a continuous linear regression.	The Sequential Bayes Factor with maximal n" (SBF+maxN) approach (Schönbrodt & Wagenmakers, 2018) was applied to collect data in an efficient way. The minimal sample size was 500 participants, and more participants were sequentially recruited in steps of 50 until the optional stopping criterion or the maximal sample size was reached. The maximal sample size was determined by	1. Regression of dRT on continuous number magnitude (1, 2, 3, 4, 6, 7, 8, 9; see models MC-1 and PJ-1) for each task separately per participant (as in Fias et al., 1996) 2. Two two-sided Bayesian one-sample *t*-tests of SNARC slopes against zero for each task separately	This replication check aimed at validating the experimental manipulation and method applied in the current study. That is, finding the SNARC effect in both tasks by using the typical analysis was a positive control in this study, and was considered as a prerequisite for all further hypothesis tests. The sensitivity of the tests, however, depended on the final sample size determined by the SBF+maxN approach used	Because evidence was found for the continuous SNARC slopes to differ from zero and to be negative, the SNARC effect was detectable with the standard analysis, which was in line with previous literature and laid the groundwork for further hypothesis tests. If evidence had been found against the SNARC slopes to differ from zero, no SNARC effect would have been observable,	The SNARC effect is usually detected with the standard analysis in both MC and PJ. We strongly expected to find it in this study as well, especially with our large sample size. Not finding the continuous SNARC effect would have spoken against its robustness in MC and/or PJ and would have been very surprising. Note that this replication check was used as a basis for all further analyses	Strong evidence for Replication Check 1: MC-SNARC: slope = -4.03 ms, *SD* = 13.7 ms, *d* = -0.29, BF_{10} = $4.04 * 10^{9}$ PJ-SNARC: slope = -3.59 ms, *SD* = 6.37 ms, *d* = -0.56, BF_{10} = $1.02 * 10^{35}$

		drawing 5000 simulated datasets around the effect size of interest (Cohen's *d* = 0.2) to estimate the probability to obtain evidence for or against our hypotheses in the Bayesian framework (analogously to statistical power simulations in the frequentist framework). The respective Bayesian test was conducted for the dataset in each of these 5000 simulations. Specifically, the required sample size was determined by making sure that the proportion of Bayesian tests revealing at least moderate evidence for the alternative hypothesis (BF10 > 3) or null hypothesis (BF10 < 1/3) is .90. This		for the hypotheses.	which would have been highly unlikely. If evidence had been found for the continuous SNARC slopes to differ from zero and to be positive, a reversed SNARC effect would have been observable (i.e., association of small/large numbers with the right/left, respectively), which would have been highly unlikely.	(i.e., finding the SNARC effect with the standard analysis in both tasks was a prerequisite for testing the hypotheses in this study).	
Can the presence of a MARC effect be replicated in PJ, and can its absence be replicated in MC?	**Replication Check 2:** A MARC effect arises in PJ, but not in MC, because the activation of parity seems not to be automatic when parity is task-irrelevant.		1. Regression of dRT on contrast-coded number parity (i.e., -0.5 for odd and +0.5 for even numbers; see models MC-3 and PJ-3) for each task separately per participant (as in Nuerk et al., 2004) 2. Two two-sided Bayesian one-	Replication checks aimed at replicating observations that are typically made in the used paradigm, instead of testing new hypotheses. The sensitivity of the tests, however, depended on the final sample size determined by the SBF+maxN approach used	If evidence had been found for the MARC slopes to differ from zero and to be negative, the MARC effect would have been detectable. This would have been in line with previous literature for PJ. Because evidence was found against the	The MARC effect is usually detected in PJ, but not in MC. A theory accounting for this is that the spatial mapping of number parity is automatic, but only when number parity is activated by the task instructions. However, number parity is not activated	Strong evidence against Replication Check 2: MC-MARC: slope = -0.76 ms, *SD* = 23.6 ms, *d* = -0.03, BF_{10} = 0.06 PJ-MARC: slope = -1.55 ms, *SD* = 89.38 ms, *d* = -0.02, BF_{10} = 0.05

		procedure resulted in a maximal sample size of n = 1700, see RMarkdown script at https://osf.io/4wpv6/.	sample *t*-tests of MARC slopes against zero for each task separately	for the hypotheses.	MARC slopes to differ from zero, no MARC effect was observable. This was in line with previous literature for MC. If evidence had been found for the MARC slopes to differ from zero and to be positive, a reversed MARC effect would have been observable (i.e., association of odd/even numbers with the right/left, respectively), which would have been highly unlikely.	when being task-irrelevant, thus no spatial mapping occurs for it. We expected a replication in this study as well, and not finding the MARC effect was rather surprising in a large Western sample. However, because a considerable proportion of Western individuals reveals a reversed MARC effect (e.g., descriptively 60% regular and 40% reversed in Cipora, Soltanlou, et al., 2019), not finding evidence for the regular MARC effect did not preclude further analyses.	
Can responses be replicated to be faster in MC than in PJ?	**Replication Check 3:** RT are shorter in MC than in PJ.		One two-sided Bayesian paired *t*-test to compare mean RT per participant	Replication checks aimed at replicating observations that are typically	Evidence for faster responses in MC than in PJ was in line with	The processing of number magnitude is highly automatized and	Strong evidence for Replication Check 3:

			between tasks against zero	made in the used paradigm, instead of testing new hypotheses. The sensitivity of the tests, however, depended on the final sample size determined by the SBF+maxN approach used for the hypotheses.	previous literature.	a primitive in numerical cognition (Tzelgov et al., 2015). In contrast, the processing of number parity is not as highly automatized; it needs to be executed intentionally and is therefore slower. Finding no difference in RT between tasks or even the reversed pattern would have been highly surprising.	$M_{RT\ MC\text{-}PJ}$ = -50.31 ms, d = -1.04, BF_{10} = 2.46 * 10^{95}
Can the Numerical Distance Effect in MC be replicated?	**Replication Check 4:** A Numerical Distance Effect arises in MC (i.e., faster reactions with increasing numerical distance between the stimulus and the reference number 5).		1. Regression of RT on numerical distance (i.e., difference between the number and the criterion number 5) and continuous magnitude (1, 2, 3, 4, 6, 7, 8, or 9) for each participant separately (as in Hohol et al., 2020)	Replication checks aimed at replicating observations that are typically made in the used paradigm, instead of testing new hypotheses. The sensitivity of the tests, however, depended on the final sample size determined by the SBF+maxN approach used	Because evidence was found for the slopes to differ from zero and to be negative, the Numerical Distance Effect was detected, which was in line with previous literature. If evidence had been found against the slopes to differ from zero, no	The Numerical Distance Effect is usually detected in MC. We strongly expected to find it in this study as well, especially with our large sample size. Not finding the Numerical Distance Effect would have spoken against its robustness and would have been very surprising.	Strong evidence for Replication Check 4: slope = -10.96 ms, SD = 7.30 ms, d = -1.50, BF_{10} = 1.57 * 10^{154}

			2. One two-sided Bayesian one-sample *t*-test of numerical-distance slopes against zero	for the hypotheses.	Numerical Distance Effect was observable. If evidence had been found for the slopes to differ from zero and to be positive, a reversed Numerical Distance Effect was observable (i.e., faster reactions with increasing numerical distance between the stimulus and the reference number 5), which is highly unlikely.		
What is the shape of the SNARC effect in MC?	**Hypothesis 1:** The MC-SNARC is categorical.		1. Regression of dRT on continuous magnitude (model MC-1) and on categorical (contrast-coded with -0.5 for small and +0.5 for large numbers) magnitude (model MC-2) 2. Logit-transformation of	The main goal of this study was to determine the shape of the SNARC effect in MC and PJ (Hypotheses 1 and 2). For comparing the fit of a continuous and a categorical model against each other in MC and PJ separately, the effect size of	Evidence for a higher logit-transformed R^2 for the categorical (MC-2) than continuous (MC-1) MC-SNARC spoke for a stepwise shape of the MC-SNARC (in line with Hypothesis 1).	Results from many previous studies lead to the hypothesis that the SNARC effect is categorical in MC. In this investigation with a sufficiently large sample, we investigated its shape systematically by comparing the fit of the two	Strong evidence for Hypothesis 1: Continuous model MC-1: R^2 = 0.420, R^2_{logit} = -0.325 Categorical model MC-2: R^2 = 0.468, R^2_{logit} = -0.127 Comparison: BF_{10} = 141464.40

			R^2 for each model for each participant separately to approximate two normal distributions 3. Comparison of the logit-transformed R^2 between MC-1 and MC-2 in a two-sided paired *t*-test (as in Koch et al., 2023) Additional purely Bayesian approach: Leave-one-out cross validation to determine which of the two predictors better fits the data	interest (ESOI) had to be chosen in a standardized unit. We determined Cohen's *d* = 0.2 as ESOI, because it is considered to be a small effect (Cohen, 1988). This ESOI was used to determine the maximal sample size for the SBF+maxN approach.		statistical models.	Evidence against Hypothesis 1: Leave-one-out cross validation: ELPD difference between models of -15.29 with *SE* = 6.08, indicating that MC-1 outperformed MC-2.
What is the shape of the SNARC effect in PJ?	**Hypothesis 2:** The PJ-SNARC is continuous.		1. Regression of dRT on continuous magnitude (model PJ-1) and on categorical (contrast-coded with -0.5 for small and +0.5 for large numbers) magnitude (model PJ-2)		Evidence for a higher logit-transformed R^2 for the continuous (PJ-1) than categorical (PJ-2) PJ-SNARC spoke for a linear shape of the PJ-SNARC (in line with Hypothesis 2).	Results from many previous studies lead to the hypothesis that the SNARC effect is continuous in PJ. In this investigation with a sufficiently large sample, we investigated its shape	Strong evidence for Hypothesis 2: Continuous model PJ-1: R^2 = 0.195, R^2_{logit} = -1.421 Categorical model PJ-2: R^2 = 0.164, R^2_{logit} = -1.629

			2. Logit-transformation of R^2 for each model for each participant separately to approximate two normal distributions 3. Comparison of the logit-transformed R^2 between PJ-1 and PJ-2 in a two-sided paired *t*-test (as in Koch et al., 2023) Additional purely Bayesian approach: Leave-one-out cross validation to determine which of the two predictors better fits the data			systematically by comparing the fit of the two statistical models.	Comparison: BF_{10} = 319751.90 Further evidence for Hypothesis 2: Leave-one-out cross validation: ELPD difference between models of -21.55 with *SE* = 4.92, indicating that PJ-1 outperformed PJ-2.
Can the Numerical Size Effect be found in both tasks, and does it differ between tasks regarding its strength?	**Hypothesis 3a:** A Numerical Size Effect arises in both tasks. **Hypothesis 3b:** The Numerical Size Effect is		1. Regression of RT on numerical distance (i.e., difference between the number and the criterion number 5) and continuous magnitude (1, 2, 3, 4, 6, 7, 8, or 9)	As explained above, we chose Cohen's *d* = 0.2 as ESOI for Hypotheses 1 and 2. We decided to use the same ESOI for Hypotheses 3a and 3b for	Because evidence was found for the slopes to differ from zero and to be positive, the Numerical Size Effect was detected, which was in line with	The Numerical Size Effect is usually detected in MC and PJ. We expected to find it in this study as well, especially with our large sample size. Not finding the Numerical	Moderate/strong evidence for Hypothesis 3a: MC: slope = 0.38 ms, *SD* = 3.12 ms, *d* = 0.12, BF_{10} = 4.40 PJ: slope = 1.06 ms, *SD* = 2.64

	stronger in MC than in PJ.		for each participant separately (as in Hohol et al., 2020) 2. One two-sided Bayesian one-sample *t*-test of continuous-magnitude slopes against zero 3. One two-sided Bayesian paired *t*-test between continuous-magnitude slopes	consistency reasons.	previous literature. If evidence had been found against the slopes to differ from zero, no Numerical Size Effect effect would have been observable. If evidence had been found for the slopes to differ from zero and to be negative, a reversed Numerical Size Effect would have been observable (i.e., faster reactions with increasing numerical magnitude), which would have been highly unlikely.	Size Effect would have spoken against its robustness and be surprising, although some individuals seem to reveal a reversed Numerical Size Effect (showing that it is less consistent than the Numerical Size Effect; Hohol et al., 2020).	ms, d = 0.40, BF_{10} = 2.67 * 10^{18} Strong evidence for the opposite of Hypothesis 3b: Comparison: BF_{10} = 204.69

Note. This Study Design Table was filled in prior to data collection – only the rightmost column showing the truly observed outcomes was added post data collection.

Table 2

Overview of dRT predictors

Continuous magnitude	1	2	3	4	6	7	8	9
Categorical magnitude: Boundary 5	-0.5	-0.5	-0.5	-0.5	+0.5	+0.5	+0.5	+0.5
Boundary 2.5 (Exploratory 4)	-0.5	-0.5	+0.5	+0.5	+0.5	+0.5	+0.5	+0.5
Boundary 3.5 (Exploratory 4)	-0.5	-0.5	-0.5	+0.5	+0.5	+0.5	+0.5	+0.5
Boundary 6.5 (Exploratory 4)	-0.5	-0.5	-0.5	-0.5	-0.5	+0.5	+0.5	+0.5
Boundary 7.5 (Exploratory 4)	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	+0.5	+0.5
Parity	-0.5	+0.5	-0.5	+0.5	+0.5	-0.5	+0.5	-0.5

Note. This table gives an overview of the dRT predictors that were used in the regression models summarized in Table 3. Continuous magnitude is equal to the actual presented stimulus. Categorical magnitude was contrast-coded with -0.5 for smaller vs. +0.5 for larger numbers (Boundary 5, i.e., a categorization into numbers 1 to 4 vs. 6 to 9, is used for most analyses, whereas different boundaries are used in Exploratory 4). Number parity was contrast-coded with -0.5 for odd and +0.5 for even numbers.

dRT effects (i.e., SNARC and MARC effects). For Replication Checks 1 and 2, as well as Hypotheses 1 and 2, we fitted four regression models per participant and per task (for an overview, see Table 3). First, the presence of a SNARC effect in both tasks (Replication Check 1, used as a positive control) was tested in a repeated-measures regression as usually done in SNARC research (see Fias et al., 1996, adapted from Lorch & Myers, 1990). For this, dRT were regressed on continuous

magnitude in models MC-1 and PJ-1 for each participant separately. The resulting slopes were tested against zero in a two-sided Bayesian one-sample *t*-test, with Bayesian evidence for a difference from zero indicating a continuous SNARC effect, and negative value of the slope indicating the typical SNARC effect.

Then, the presence of the MARC effect (Replication Check 2) was tested in both tasks with the same repeated-measures regression approach as the SNARC effect. For this, regression models MC-3 and PJ-3 were computed. These models contain a contrast-coded parity predictor and the magnitude predictor with the better fit in the previous test. Because the number-parity predictor and the number-magnitude predictor were orthogonal to each other (i.e., their correlation is zero), both could be concurrently included within one regression model without affecting the respective other parameter estimate. Then, the slopes were tested against zero in a two-sided Bayesian one-sample *t*-test, with evidence for a difference from zero indicating a MARC effect, which is only expected in PJ but not MC.

Next, we investigated whether the MC-SNARC and the PJ-SNARC were continuous or categorical (i.e., which number magnitude predictor fitted the observed dRT better; Hypotheses 1 and 2). For this, besides regressing dRT on continuous magnitude in models MC-1 and PJ-1 as previously described, they were regressed on categorical magnitude (contrast-coded as in Table 2) in MC-2 and PJ-2. Then, we logit-transformed the R^2 for each model for each participant separately to approximate two normal distributions and compare the logit-transformed R^2 between the two models in a two-sided Bayesian paired *t*-test (as Koch et al., 2023, did this in a frequentist approach). In Hypotheses 1 and 2, we predicted a better fit of model MC-2 compared to MC-1 and of PJ-1 compared to PJ-2. Additionally, to confirm these findings via a Bayesian approach, dRT were regressed on continuous and categorical magnitude for both PJ and MC in four separate Bayesian models, and in each task, a leave-one-out

cross validation was performed. An overview of possible SNARC effect shapes and the corresponding regression models tested in the current study can be found in Figure 1.

Table 3

Overview of regression models that were fitted for each participant

Magnitude classification	
MC-1	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous}$
MC-2	$dRT \sim \beta_0 + \beta_1 * magnitude_{categorical}$
MC-3[a]	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous/categorical} + \beta_2 * parity$
MC-4	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous} + \beta_2 * magnitude_{categorical}$
Parity judgment	
PJ-1	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous}$
PJ-2	$dRT \sim \beta_0 + \beta_1 * magnitude_{categorical}$
PJ-3[b]	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous/categorical} + \beta_2 * parity$
PJ-4	$dRT \sim \beta_0 + \beta_1 * magnitude_{continuous} + \beta_2 * magnitude_{categorical}$

Note. Four regression models were fitted for each participant separately for MC (MC-1, MC-2, MC-3, and MC-4) and PJ (PJ-1, PJ-2, PJ-3, and PJ-4). The predictors used in the models are specified in Table 2. In each model, β_1 and β_2 are the coefficients of the respective predictors for number magnitude or number parity. β_0 denotes the model intercept. [a] For MC-3, the better magnitude predictor from MC-1 and MC-2 was used. [b] For PJ-3, the better magnitude predictor from PJ-1 and PJ-2 was used.

RT effects (i.e., Numerical Size and Distance effects). For Replication Checks 3 and 4 as well as for Hypotheses 3a and 3b, RT were examined in detail. First, to test whether reactions are on average shorter in MC than in PJ (Replication

Check 3), we compared mean RT per participant between tasks in a two-sided Bayesian paired *t*-test. Next, the presence of a Numerical Distance Effect in MC (Replication Check 4) and of a Numerical Size Effect in both MC and PJ (Hypotheses 3a and 3b) was investigated with the repeated-measures regression approach (as in Hohol et al., 2020). In MC, RT were regressed on numerical distance (i.e., difference between the number and the criterion number 5) and continuous magnitude (1, 2, 3, 4, 6, 7, 8, or 9) for each participant separately. Because numerical distance and magnitude were orthogonal (i.e., their correlation is zero), both could be concurrently included within one regression model without affecting the respective other parameter estimate. In PJ, RT were only regressed on continuous magnitude for each participant separately. Next, resulting regression slopes was tested against zero in a two-sided Bayesian one-sample *t*-test for each task separately. Evidence for negative slopes for the numerical distance predictor indicated faster reactions for larger numerical distance, reflecting the Numerical Distance Effect (Replication Check 4). Evidence for positive slopes for the magnitude predictor indicated slower reactions for increasing number magnitude, reflecting the Numerical Size Effect (Hypothesis 3a). Last, we tested whether the magnitude of the Numerical Size Effect was stronger in MC than in PJ (Hypothesis 3b) by comparing resulting slopes between tasks in a two-sided Bayesian paired *t*-test.

Exploratory data analysis

Influence of task and block orders. After the analyses for replication checks and hypotheses, we investigated task-order effects on both the MC-SNARC and the PJ-SNARC (Exploratory 1). That is, we tested task-order effects by comparing SNARC slopes in Conditions 1 and 2 (first MC, second PJ) with Conditions 3 and 4 (first PJ, second MC) for each task separately (see Figure 2 for an overview of experimental conditions). The predictor that fitted better in the respective task was used (according

to Hypotheses 1 and 2). For this, we ran two two-sided Bayesian independent-samples t-tests.

Further, we exploratorily tested compatibility-order effects on the MC-SNARC by comparing SNARC slopes in Conditions 1 and 3 (first SNARC-incompatible, second SNARC-compatible) with SNARC slopes in Conditions 2 and 4 (first SNARC-compatible, second SNARC-incompatible; Exploratory 2a). Note that we used the categorical or continuous slope here depending on which of both described the MC-SNARC better (i.e., depending on the outcome regarding Hypothesis 1). Similarly, we tested compatibility-order effects on the MARC effect in PJ by comparing MARC slopes in Conditions 1 and 3 (first MARC-incompatible, second MARC-compatible) with MARC slopes in Conditions 2 and 4 (first MARC-compatible, second MARC-incompatible; Exploratory 2b). For this, we ran two two-sided Bayesian independent-samples *t*-tests. Evidence for a stronger SNARC/MARC effect in Conditions 2 and 4 compared to Conditions 1 and 3 reflected larger compatibility effects when the response-key assignment is first compatible and then incompatible, and vice versa.

Exact shape of the SNARC effect. Next, a fourth model including both continuous and categorical magnitude was fitted for both tasks (MC-4 and PJ-4). Both resulting slopes were tested against zero in two-sided Bayesian one-sample *t*-tests, with Bayesian evidence for both slopes being different from zero, indicating a mixed shape of the SNARC effect, as illustrated in Panel C in Figure 1 (Exploratory 3). This would mean that the dRT regression slope is negative within small and within large numbers, while there is a categorical step between numbers 4 and 6 (for an empirical observation of a such pattern, see Figure 2b in Nuerk, Bauer, et al., 2005).

Then, we fitted five categorical models for each participant per task. The models were analogous to MC-2 and PJ-2. They differed regarding the boundary between “small” and “large” numbers (contrast-coded with -0.5 and +0.5, respectively).

Specifically, we ran five regression models while classifying 1 to 2 vs. 3 to 9, 1 to 3 vs. 4 to 9, 1 to 4 vs. 6 to 9, 1 to 6 vs. 7 to 9, and 1 to 7 vs. 8 to 9 as “small” and “large” numbers (see Table 2). Note that 1 to 4 vs. 6 to 9 corresponds to the MC-2 or PJ-2 (used to test Hypotheses 1 and 2). For each participant, we determined the most likely underlying categorization by descriptively comparing, which of the five models had the best fit to the data in terms of R^2. We present the distribution of favored categorical models across the sample for each task in the Results section. Next, we logit-transformed the R^2 for the favored categorical model and for the continuous model for each participant separately and compared the logit-transformed R^2 between the two models in a two-sided Bayesian paired *t*-test (Exploratory 4). Moreover, we regressed dRT on continuous and categorical magnitude (using the favored categorical predictor for each participant) for both PJ and MC in Bayesian models and perform a leave-one-out cross validation (as for Hypotheses 1 and 2).

As a robustness check, to find out whether the determined boundary between “small” and “large” numbers was reliable for each participant and not only due to random measurement noise, we determined its split-half reliability by splitting valid trials with the odd-even method based on presentation order (i.e., 1^{st}, 3^{rd}, 5^{th}, etc. trial vs. 2^{nd}, 4^{th}, 6^{th}, etc. trial). Subsequently, we again computed five models per participant and per task, separately for each half of all valid trials, and determined which of the five models has the largest R^2 in each half of the experiment. For each half per participant within each task, we coded the boundary of the model with the highest R^2 with the cardinal boundary values 2.5, 3.5, 5, 6.5, or 7.5. We then computed the Pearson product-moment correlation between the favored categorical models across halves within each task (note that coding the favorable model ordinally with 1, 2, 3, 4, and 5 to calculate the Spearman rank correlation would lead to almost the same results, as the correlation between the cardinal and ordinal values is $r = .997$). Next,

we applied the Spearman-Brown correction to each correlation to adjust for task length and descriptively evaluate the correlations in terms of whether the boundaries between "small" and "large" numbers are related between both halves. We also created alluvial plots showing the stability of the boundaries within participants across halves for each task separately. Finally, we computed the split-half reliability for the basic categorical MC-SNARC (using the classification of 1 to 4 vs. 6 to 9 as "small" vs. "large") and continuous PJ-SNARC. These results were taken into account when interpreting the split-half reliability of the favored categorical model (as a part of Exploratory 4). Note that, to our knowledge, the split-half reliability has never been reported for the categorical MC-SNARC. For the split-half reliability of the continuous PJ-SNARC, values between .43 and .82 have been reported in the literature (for an overview, see Cipora, van Dijck, et al., 2019; Table 1).

Time course of the SNARC effect. Moreover, we explored whether the shape of the SNARC effect differed between earlier and later phases within each task (Exploratory 5). Importantly, it was not possible to determine the SNARC effect in the first or second block of each task separately, because both blocks are needed in order to calculate the differences between left- and right-hand responses. Therefore, we computed the models MC-4 and PJ-4 and tested the resulting slopes for both a continuous predictor and a categorial predictor against zero in two-sided Bayesian one-sample *t*-tests, but instead of considering all 30 repetitions per block, we only considered the first or second halves of both blocks within each task (i.e., first or second 15 repetitions of each number in one and in the other response-to-key assignment). This way, we could investigate whether early trials in each response-to-key assignment lead to a different SNARC shape than late trials.

Relations between the MC-SNARC and the PJ-SNARC. Further, we tested whether the MC-SNARC is stronger than the PJ-SNARC by comparing the slopes for

continuous magnitude predictors resulting from models MC-1 and PJ-1 in a two-sided Bayesian paired *t*-test (Exploratory 6). Lastly, we calculated Pearson's correlation between the MC-SNARC and PJ-SNARC slopes (Exploratory 7). For this, the predictor leading to a better model fit per task across all participants was used (categorical in MC according to Hypothesis 1 and linear in PJ according to Hypothesis 2). We ran a two-sided Bayesian Pearson correlation test to see whether the spatial mapping of number magnitude within participants was similar in both tasks.

Data quality and positive controls

To control the data quality in our study, we implemented a seriousness check (Reips, 2009) as well as a self-assessment of noise, distractions, and other difficulties. To make sure that we only analyzed trials that reflected mental processes in correctly executed MC or PJ, we only included correctly answered trials, trimmed RT, and only included data from participants with a minimum of 75% remaining valid trials (as described in the data preprocessing pipeline). Moreover, the test of the MC-SNARC and PJ-SNARC analyzed with the traditional linear regression (Replication Check 1) served as positive control. Importantly, we considered this positive control as a prerequisite for all further analyses and only proceeded with testing the other hypotheses, because we could find at least moderate Bayesian evidence for both the continuous MC-SNARC and the continuous PJ-SNARC at the group level. Finally, in additional replication checks, we aimed to replicate results from previous studies to validate our investigation.

Possible limitations and unexpected outcomes

Importantly, including both continuous and categorical magnitude within one single regression model (as in MC-4 and PJ-4) is problematic because of collinearity between these two predictors (i.e., $r = .913$). However, we still decided to compute one such regression model for each task because the true shape of the SNARC effect

might be determined by both continuous and categorical number magnitude simultaneously.

In the present study, we tested the two most frequently used versions of MC and PJ (i.e., with symbolic single-digit numbers) in a sample in which the SNARC effect is not controversial (i.e., Western culture with left-to-right reading and writing direction). Future studies will show whether our results hold true for different types of stimuli and for different samples.

Stage-2 Registered Report

Results

The data collection and the data analyses were conducted as described in the peer-reviewed Stage-1 Registered Report accepted by the Peer Community In. Additional non-preregistered analyses can be found in Appendix B. All R scripts for data preprocessing and analysis as well as all anonymized datasets can be found at https://osf.io/g48s2/. The following R packages were used: *BayesFactor* by Morey et al. (2015), *brms* by Buerkner (2017), *data.table* by Barrett et al. (2024), *dplyr* by Wickham, François et al. (2023), *ggalluvial* by Brunson & Read (2023), *ggplot2* by Wickham et al. (2024), *ggpubr* by Kassambara (2025), *loo* by Vehtari et al. (2017), *scales* by Wickham et al. (2025), and *tidyr* by Wickham, Vaughan, et al. (2023). Observed outcomes were added to the Study Design Table (see rightmost column of Table 1) that was filled in prior to data collection.

Data preprocessing

Our final sample recruited in line with the SBF+maxN approach consisted of 701 participants who completed both tasks. Out of these participants, none were under 18 and three were over 40. Two indicated they only wanted to have a look at the

experiment but not seriously participate. Further, one participant reported a very/extremely noisy environment and two reported multiple major distractions. Thirty-one participants did not use their left/right index finger for the left/right response key consistently as described in the instructions. In total, the data of 39 participants were therefore excluded from the analysis in a first preprocessing step. Next, after discarding practice trials, exclusion criteria were applied at the trial level for each task separately:

- MC: 5.05% of the trials were excluded due to incorrect responses, 0.30% due to responses faster than 200 ms, 2.23% due to responses slower than 1500 ms, and 7.13% in the iterative trimming procedure.
- PJ: 5.95% of the trials were excluded due to incorrect responses, 0.20% due to responses faster than 200 ms, 2.96% due to responses slower than 1500 ms, and 7.09% in the iterative trimming procedure.

Finally, after these exclusions at the trial level, 13 participants had at least one empty experimental cell, and 53 participants had less than 75% valid trials remaining (31 in MC and 33 in PJ, of which 11 overlappingly in both tasks). Their data were thus excluded from analyses, and data from 609 participants remained for analysis. Originally, we had planned to apply dropout analysis (Reips, Overlander & Bannert, 2025) to detect possible motivational confounding (Reips, 2000, 2021), but given there were only few dropouts, such an analysis was not necessary.

Descriptive self-reported information about the participants is summarized in Table 4. Most of them used the default response keys, but 38 replaced D and 41 replaced K (among which 35 replaced both keys). Parameter estimates for the four main regression models (see Table 3) can be found in Table 6.

Table 4

Descriptive self-reported information about the sample (N = 609 after exclusions)

Demographic item	**Answer**	
Age in years	Mean (SD)	30.0 (5.61)
Gender	Woman	297 (48.8%)
	Man	309 (50.7%)
	Other	3 (0.5%)
Native language	English	608 (99.8%)
	Others	1 (0.2%)
Handedness	Right-handed	530 (87.0%)
	Left-handed	56 (9.2%)
	Ambidextrous	23 (3.8%)
Finger counting habits	Right-Starters	338 (55.5%)
	Left-Starters	219 (36.0%)
	Does not know or no preference	52 (8.5%)
Finger counting stability	Always	275 (45.2%)
	Mostly	240 (39.4%)
	Slightly more often	38 (6.2%)
	Does not know or no preference	2 (0.3%)
	No answer	54 (8.9%)

Table 6

Summary of the predictors in both tasks (i.e., mean and SD parameter estimates) using models in Table 3

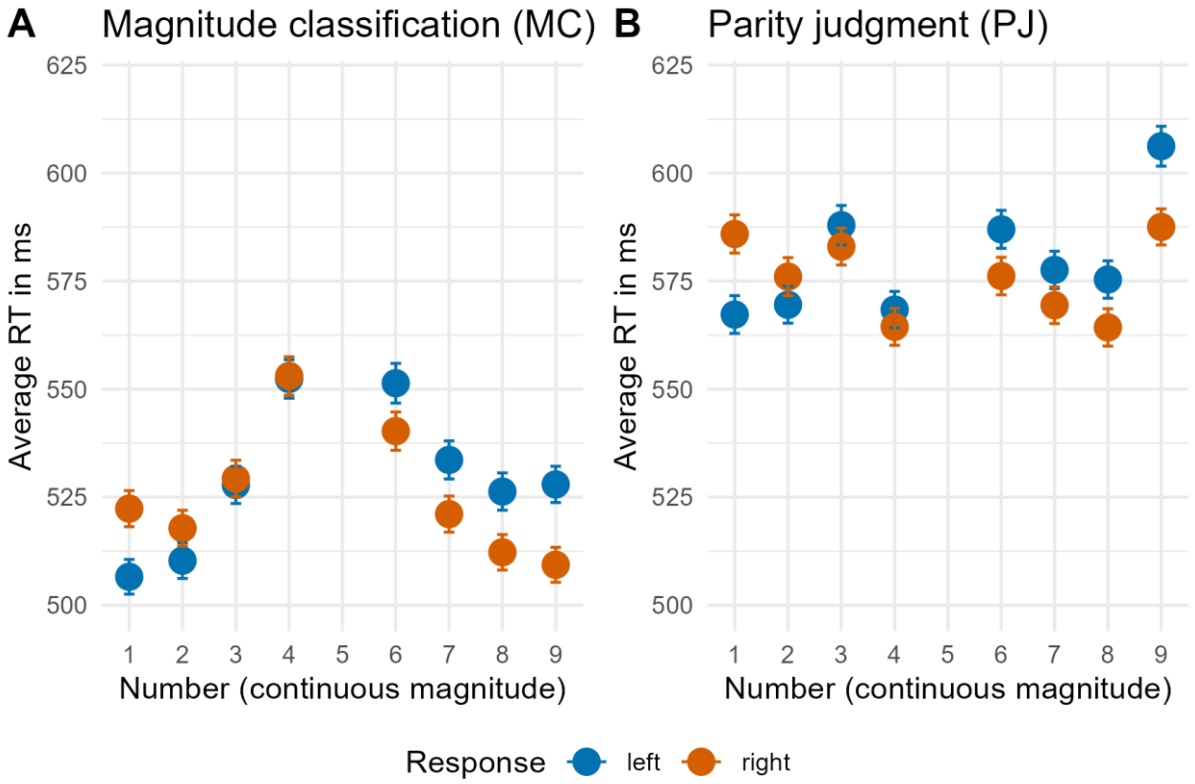
MC			PJ		
Model	**Predictor**	**Mean slope (SD)**	**Model**	**Predictor**	**Mean slope (SD)**
MC-1	β_1	-4.03 (13.70)	PJ-1	β_1	-3.59 (6.37)
MC-2	β_1	-20.44 (81.46)	PJ-2	β_1	-16.25 (32.22)
MC-3	β_1	-13.47 (76.50)	PJ-3	β_1	-7.75 (22.35)
	β_2	-0.76 (23.60)		β_2	-1.55 (89.38)
MC-4	β_1	-3.77 (10.35)	PJ-4	β_1	-5.29 (12.29)
	β_2	-1.59 (96.23)		β_2	10.21 (58.84)

Confirmatory data analysis

The average RT for all included trials was 527 ms (*SD* = 100 ms) in MC and 577 ms (*SD* = 101 ms) in PJ. Figure 3 illustrates the mean RT including their standard error (SE) per stimulus and per response in each task.

Figure 3

Mean RT (error bars represent ± 1 SE) in MC (Panel A) and PJ (Panel B) per digit and per response side for trials included in the data analysis



dRT effects (i.e., SNARC and MARC effects). ***Replication Check 1.*** Bayesian one-sample *t*-tests of the continuous magnitude slopes against zero revealed strong evidence for an MC-SNARC (slope = -4.03 ms, *SD* = 13.70 ms, *d* = -0.29, BF_{10} = 4.04 * 10^{9}), and strong evidence for a PJ-SNARC (slope = -3.59 ms, *SD* = 6.37 ms, *d* = -0.56, BF_{10} = 1.02 * 10^{35}). The positive control was thus successful, which is in line with our expectation and fulfills the prerequisite for testing all further hypotheses. The MC-SNARC is plotted in Figure 4, and the PJ-SNARC is plotted in Figure 5.

Replication Check 2. Bayesian one-sample *t*-tests of the contrast-coded parity slopes against zero revealed strong evidence against an MC-MARC (slope = -0.76 ms, *SD* = 23.60 ms, *d* = -0.03, BF_{10} = 0.06), as well as against a PJ-MARC (slope = -1.55

ms, SD = 89.38 ms, d = -0.02, BF_{10} = 0.05). Note that this is at odds with several studies demonstrating the MARC effect in PJ (i.e., no replication).

Hypothesis 1. A two-sided Bayesian paired *t*-test revealed strong evidence for a better fit of the categorical model MC-2 (R^2 = 0.468, R^2_{logit} = -0.127) than of the continuous model MC-1 (R^2 = 0.420, R^2_{logit} = -0.325) with BF_{10} = 141464.40. However, a the ELPD difference between models (MC-1 vs. MC-2) obtained in the leave-one-out cross validation was -15.29 with SE of 6.08, indicating that MC-1 outperformed MC-2. In total, results regarding Hypothesis 1 were mixed.

Hypothesis 2. A two-sided Bayesian paired *t*-test revealed strong evidence for a better fit of the continuous model PJ-1 (R^2 = 0.195, R^2_{logit} = -1.421) than of the categorical model PJ-2 (R^2 = 0.164, R^2_{logit} = -1.629) with BF_{10} = 319751.90. In line with our hypothesis, this result indicates a rather continuous than categorical shape of the PJ-SNARC. This was further supported by an ELPD difference (PJ-1 vs. PJ-2) of -21.55 with a SE of 4.92 obtained in the leave-one-out cross validation, indicating that PJ-1 outperformed PJ-2.

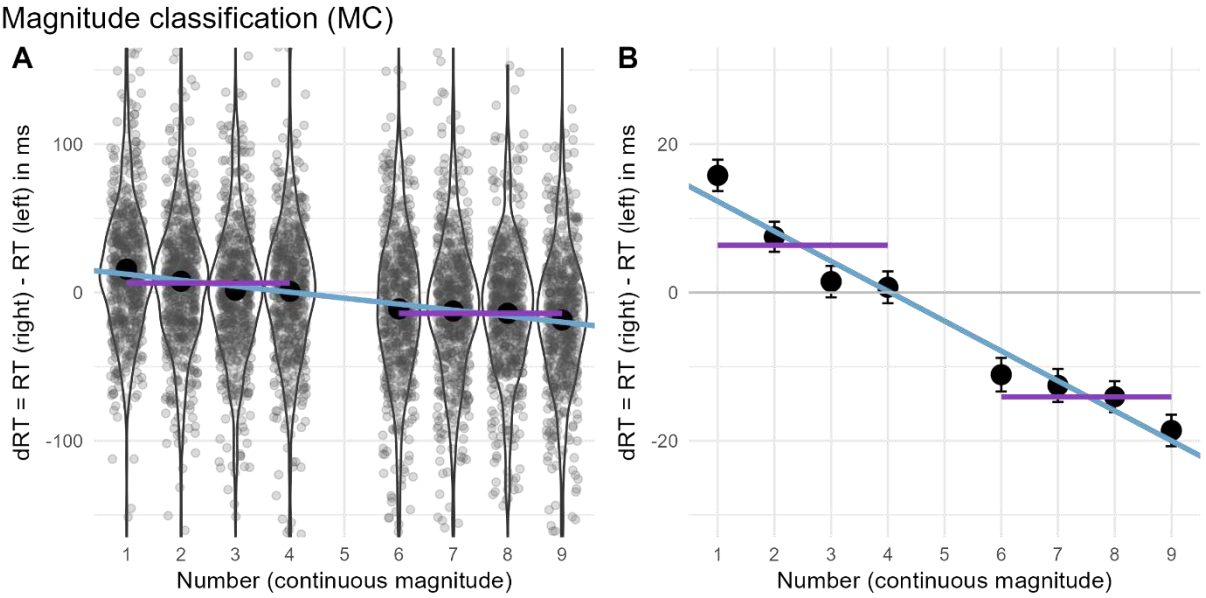
RT effects (i.e., Numerical Size and Distance effects). ***Replication Check 3.*** A two-sided Bayesian paired *t*-test comparing mean RT averaged across all trials within each participant revealed strong evidence for shorter RT in MC than in PJ ($M_{RT\ MC\text{-}PJ}$ = -50.31 ms, d = -1.04, BF_{10} = 2.46 * 10^{95}). This result is in line with our expectation and previously reported results (see Figure 3).

Replication Check 4. A two-sided Bayesian one-sample *t*-test of the numerical distance predictor for RT against zero provided strong evidence for a Numerical Distance Effect in MC (slope = -10.96 ms, SD = 7.30 ms, d = -1.50, BF_{10} = 1.57 * 10^{154}). This replicates faster responses in MC for larger numerical distances (see left panel of Figure 3).

Hypotheses 3a and 3b. Two-sided Bayesian one-sample t-tests of the continuous magnitude predictor for RT against zero provided moderate evidence for a Numerical Size Effect in MC (slope = 0.38 ms, *SD* = 3.12 ms, *d* = 0.12, BF_{10} = 4.40) and strong evidence for a Numerical Size Effect in PJ (slope = 1.06 ms, *SD* = 2.64 ms, *d* = 0.40, $BF_{10} = 2.67 * 10^{18}$). This is in line with faster responses for smaller than larger number magnitudes in both tasks, as predicted in Hypothesis 3a and as previously reported in the literature. However, a two-sided Bayesian paired *t*-test revealed strong evidence for a stronger Numerical Size Effect in PJ than in MC (BF_{10} = 204.69), which is the opposite of what was expected in Hypothesis 3b.

Figure 4

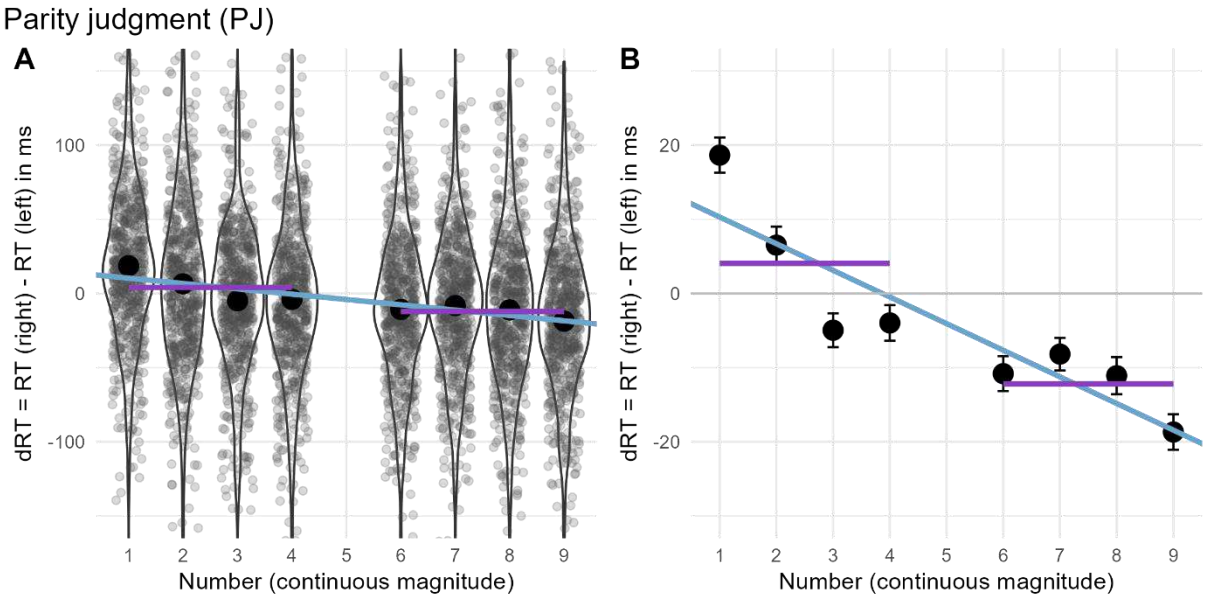
The SNARC effect in MC



Note. Panels A and B both illustrate the same data from MC, with large black dots representing mean dRT per number magnitude across all participants. The blue line with the negative slope corresponds to the continuous model MC-1, the purple horizontal lines correspond to the categorical model MC-2. In Panel A, smaller gray dots summarized by violins represent mean dRT per number magnitude per participant separately. In Panel B, error bars represent ± 1 SE. Please note that the y-axis scaling differs between Panels A and B.

Figure 5

The SNARC effect in PJ



Note. Panels A and B both illustrate the same data from PJ, with large black dots representing mean dRT per number magnitude across all participants. The blue line with the negative slope corresponds to the continuous model PJ-1, the purple horizontal lines correspond to the categorical model PJ-2. In Panel A, smaller gray dots summarized by violins represent mean dRT per number magnitude per participant separately. In Panel B, error bars represent ± 1 SE. Please note that the y-axis scaling differs between Panels A and B.

Exploratory data analysis

Influence of task and block orders. ***Exploratory 1.*** Two-sided Bayesian independent-samples *t*-tests revealed strong evidence against task-order effects both on the categorical MC-SNARC (BF_{10} = 0.09; slope = -20.89 ms in Conditions 1 and 2, and slope = -20.01 ms Conditions 3 and 4), and on the continuous PJ-SNARC (BF_{10} = 0.18; slope = -3.28 ms in Conditions 1 and 2, and slope = -3.88 ms Conditions 3 and 4).

Exploratory 2. Two-sided Bayesian independent-samples *t*-tests revealed strong evidence for a compatibility-order effect on the categorical MC-SNARC (BF_{10} =

6279.23; slope = -5.00 ms in Conditions 2 and 4, where participants started with the SNARC-compatible block, and slope = -36.23 ms in Conditions 1 and 3, where participants started with the SNARC-incompatible block). Evidence regarding a compatibility-order effect on the continuous PJ-MARC was inconclusive (BF_{10} = 0.47; slope = 5.01 ms when the first condition was MARC-compatible and slope = -8.26 ms when the first condition was MARC-incompatible).

Exact shape of the SNARC effect. ***Exploratory 3.*** The continuous and categorical SNARC slopes resulting from a joint model using both predictors (MC-4 or PJ-4) were tested against 0 separately in two-sided Bayesian one-sample *t*-tests. Evidence was strong for a continuous proportion (BF_{10} = 1.08 * 10^{15}) and against a categorical proportion (BF_{10} = 0.05) of the MC-SNARC. Evidence was strong for both a continuous (BF_{10} = 1.14 * 10^{21}) and a categorical proportion (BF_{10} = 367.30) of the PJ-SNARC. Naturally, the fits were higher for MC-4 (R^2 = 0.571, R^2_{logit} = 0.285) and PJ-4 (R^2 = 0.279, R^2_{logit} = -0.949) than for MC-2 and PJ-1, which only included the more suitable magnitude predictor but not both. These results are not in line with the categorical MC-SNARC observed in Hypothesis 1 and point towards a mixed shape of the PJ-SNARC (see Panel C in Figure 1). This might be due to high collinearity between predictors reducing the explained variance attributable to each separately, complicating interpretation. Additionally, the model is limited by the number of data points (i.e., eight stimuli) being small relative to the number of parameters estimated (i.e., two magnitude predictors). To overcome these limitations, we additionally ran Exploratory 17 (see Appendix). The reasoning was that, on the one hand, the MC-SNARC seems to follow a step-wise pattern more than a linear one (Hypothesis 1). On the other hand, as descriptively observable in Figure 4 and as revealed by a similarly high fit of the continuous model, there might still be a linear trend within small and large

numbers. We therefore tested continuous regression slopes within small and large numbers separately in Exploratory 17.

Exploratory 4. The distribution of the most likely underlying categorization into "small" and "large" numbers obtained by comparing five models with different boundaries (2.5, 3.5, 5, 6.5, or 7.5; see Table 2) per participant in each task is summarized in Table 5. Naturally, in MC, stronger evidence was found for the fit of the favored categorical (R^2 = 0.573, R^2_{logit} = -0.293) than continuous model MC-1 (R^2 = 0.420, R^2_{logit} = -0.325) with BF_{10} = 1.30 * 10^{85}. The ELPD difference (MC-1 vs. favored categorical model) obtained in the leave-one-out cross validation was -0.91 with a SE of 7.27. This difference, being relatively small as compared to the SE, fails to indicate a clear advantage for either model (with a descriptive tendency towards a better fit of the continuous model). Interestingly, in PJ, strong evidence was found for a better fit of the favored categorical model (R^2 = 0.335, R^2_{logit} = -0.685) than for the continuous model PJ-1 (R^2 = 0.195, R^2_{logit} = -1.421) with BF_{10} = 6.10 * 10^{92} as well, although the results for Hypothesis 2 revealed a better fit of the continuous model when compared to the standard categorical model. In line with this, the ELPD difference (PJ-1 vs. favored categorical model) obtained in the leave-one-out cross validation was -31.94 with a SE of 8.51, also indicating a better fit of the favored categorical model. Evidence regarding the split-half reliability of the favored categorical boundaries was inconclusive in MC (Spearman-Brown corrected estimate r = 0.16; BF_{10} = 0.97) and strongly against a difference from 0 in PJ (Spearman-Brown corrected estimate r = 0.02; BF_{10} < 0.10), as calculated with the odd-even method. Thus, categorical boundaries between "small" and "large" seemed rather unreliable, as illustrated in Figure 6. In contrast, evidence for the split-half reliability of the SNARC effect in each task to differ from 0 was strong: The odd-even method revealed r = 0.89 for the standard categorical MC-SNARC (Spearman-Brown corrected estimate r = 0.94;

BF_{10} = 1.15 * 10^{208}) and r = 0.27 for the continuous PJ-SNARC (Spearman-Brown corrected estimate r = 0.43; BF_{10} = 1.57 * 10^{9}). The higher split-half reliability for the MC-SNARC than for the PJ-SNARC is visible in Figure 7.

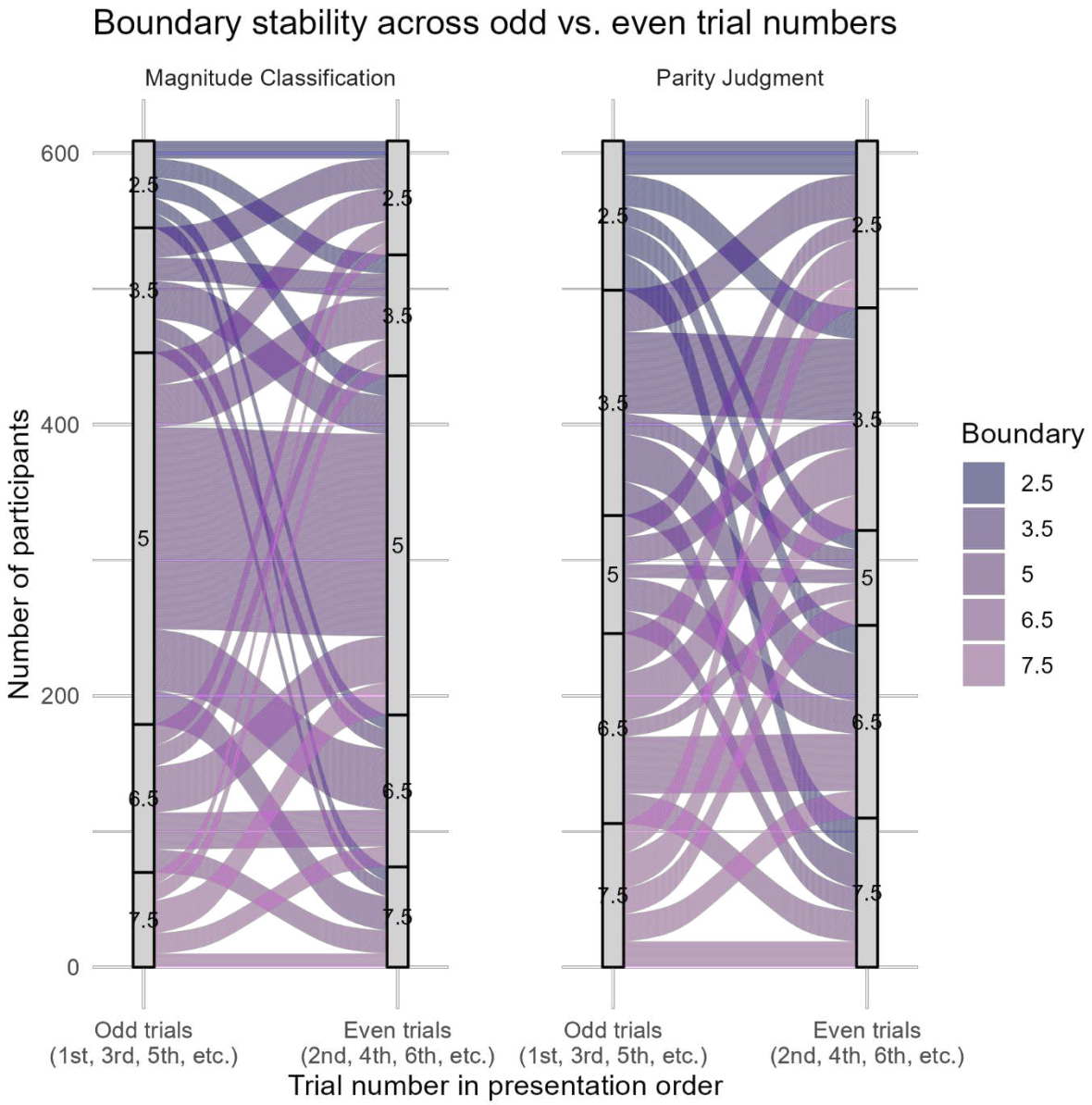
Table 5

Number and proportion of participants for each favored categorical boundary between "small" and "large" (2.5, 3.5, 5, 6.5, or 7.5) in each task (Exploratory 4).

	Number of participants (proportion)	
Favored categorical model	MC	PJ
Boundary 2.5	68 (11.2 %)	112 (18.4 %)
Boundary 3.5	87 (14.3 %)	175 (28.7 %)
Boundary 5	304 (49.9 %)	73 (12.0 %)
Boundary 6.5	100 (16.4 %)	160 (26.3 %)
Boundary 7.5	50 (8.2 %)	89 (14.6 %)

Figure 6

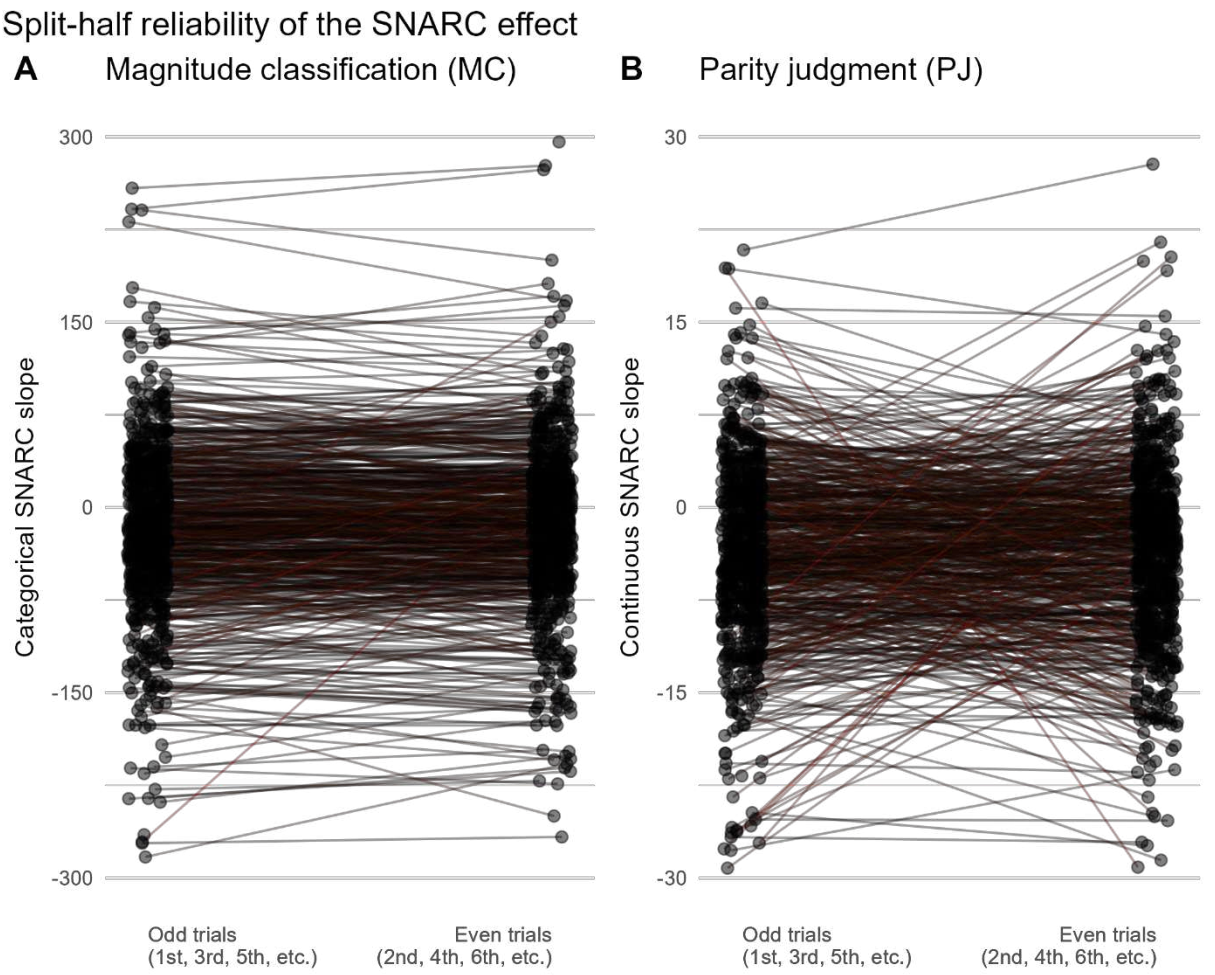
Alluvial plots illustrating the split-half reliability of categorical boundaries between "small" and "large" numbers per task (Exploratory 4)



Note. These alluvial plots show the proportion of participants for which the favored categorical boundary between "small" and "large" was 2.5, 3.5, 5, 6.5, or 7.5 in each half of each task (left panel: MC, right panel: PJ) when splitting trials based on presentation order (odd: 1st, 3rd, 5th, etc. vs. even: 2nd, 4th, 6th, etc.).

Figure 7

Plots illustrating the split-half reliability of the categorical MC-SNARC in the left panel and the continuous PJ-SNARC in the right panel (Exploratory 4)



Note. These plots illustrate the SNARC slopes for each participant as black points when calculated separately for odd and even trial numbers in the presentation order (i.e., 1st, 3rd, 5th, etc. trial vs. 2nd, 4th, 6th, etc. trial), reflecting the split-half reliability as by the odd-even method. The better fitting model according to results regarding Hypotheses 1 and 2 was used in each task (categorical MC-SNARC in the left panel and continuous PJ-SNARC in the right panel). Flatter lines connecting each participant's SNARC slopes between trial types (more black) speak for higher split-half reliability, whereas more diagonal ones (more red) speak for lower split-half reliability (with the color gradient being adjusted depending on relative steepness of the lines considering the range from -300 to 300 in MC and -30 to 30 in PJ).

Time course of the SNARC effect. ***Exploratory 5.*** Next, we considered the first and last 15 repetitions of each number within each response-to-key assignment by recomputing separate MC-4 and PJ-4 models and testing predictors in two-sided Bayesian one-sample *t*-tests. Strikingly, evidence was strong for a continuous proportion of the MC-SNARC in both the first (BF_{10} = 150792493) and last (BF_{10} = 47543464) 15 repetitions, and strongly against a categorical proportion of the MC-SNARC in both the first (BF_{10} = 0.06) and last (BF_{10} = 0.05) 15 repetitions, which is not in line with results for Hypothesis 1. Likewise, evidence was strong for a continuous proportion of the PJ-SNARC in both the first (BF_{10} = 8.75 * 10^{13}) and last (BF_{10} = 1.11 * 10^{9}) 15 repetitions. Evidence for a categorical proportion of the PJ-SNARC was inconclusive in the first (BF_{10} = 2.68) and moderately in favor in the last (BF_{10} = 5.22) 15 repetitions. To sum up, the shape of the SNARC effect seems to be similar in earlier and later phases within each task.

Relations between the MC-SNARC and the PJ-SNARC. ***Exploratory 6.*** A two-sided Bayesian paired *t*-test revealed strong evidence against a difference in strength of the continuous MC-SNARC (slope = -4.03 ms) and PJ-SNARC (slope = -3.59 ms) with BF_{10} = 0.06.

Exploratory 7. Pearson's correlation between the MC-SNARC and the PJ-SNARC (using the better predictor for each task across participants as according to Hypotheses 1 and 2, i.e., categorical in MC and continuous in PJ) was r = 0.16. A two-sided Bayesian Pearson correlation test (using the default r = 0.333 as a prior) provided strong evidence for this positive correlation with BF_{10} = 206.96.

Discussion

The present study aimed at replicating the SNARC effect, which is the hallmark effect in the field of numerical cognition, in large-scale online studies and at comparing

the two tasks most commonly used to assess the SNARC effect. Although the SNARC effect is highly robust and generalizable across many settings, at least when running the standard continuous linear regression and averaging across all participants, it also depends on the experimental situation and task (Cipora, Patro, & Nuerk, 2018). The most important goal was to compare the shape of the SNARC effect between MC, where participants decide whether numbers are smaller or larger than 5, and PJ, where participants make judgments about them being odd or even. Moreover, a within-subjects design with high statistical power allowed for testing multiple other effects and task differences, as well as the correlation of SNARC effects observed in these two tasks. In the following, findings regarding the replication checks will be summarized and discussed first, results regarding our three main hypotheses next, and observations from exploratory tests thereafter. Lastly, we give a summary and draw a conclusion with practical and theoretical implications.

Findings obtained from the replication checks

Most observations from previous studies could be replicated: We found a continuous SNARC effect in both MC and PJ (Replication Check 1; for a comparison, see Exploratory 6), responses were faster in MC than in PJ (Replication Check 3), and there was a Numerical Distance Effect in MC (Replication Check 4). However, the data revealed strong evidence against the MARC effect in both tasks. Typically, the MC-MARC is not observed (Cipora et al., 2014; Deng et al., 2018), because the processing of parity and its relation to space is not as robust as that of magnitude (Roth et al., 2025a). In other words, the SNARC effect has been shown to be more automatic than the MARC effect. It is therefore expected not to find a MARC effect in MC, where the instructions do not address parity at all. However, it is surprising that we could not detect the PJ-MARC. Notably, the MARC effect was originally found in tasks with more emphasis on language, so it is more robust when using number words instead of Arabic

numbers (Nuerk et al., 2004). Although in English, the term “odd” is not the literal negation of the term “even”, so that linguistic markedness of parity might be weak, no difference in the PJ-MARC was found in an earlier comparison between languages with and without this negation (Cipora, Soltanlou, et al., 2019).

Findings obtained from the confirmatory analysis

The main aim of the current study was to investigate the shape of the SNARC effect in the two most common tasks. As suggested in the existing literature (e.g., Wood et al., 2008), categorical magnitude in terms of numbers 1 to 4 being classified as “small” and numbers 6 to 9 being classified as “large” was a better predictor for the MC-SNARC than continuous magnitude from 1 to 9 (Hypothesis 1). However, this seems not to be the whole story, because datapoints revealed a negative slope instead of a horizontal alignment within small and large numbers when tested separately (Exploratory 17). This points towards an overall categorical MC-SNARC with a continuous proportion both in small and large numbers (i.e., the truly observed pattern visible in Panel B in Figure 4 looks most like the illustration in Panel C in Figure 1). Note that this has previously not reported for the MC-SNARC in the literature. For instance, when summarizing studies included in an early meta-analysis, the MC-SNARC looked purely categorical with horizontally arranged datapoints within small and large numbers (see Figure 1 in Wood et al., 2008). One of our own studies (Bulut et al., 2025), however, revealed a categorical step in dRT between small and large numbers with a seemingly negative slope in small but not large numbers in all three samples. It is thus surprising that the present study speaks for an overall categorical shape with continuous aspects in both small and large numbers. Regarding the PJ-SNARC, as expected from the existing literature, the continuous predictor revealed a better fit than the categorical one (Hypothesis 2). The findings are in line with a fine-graded mental representation in PJ, where magnitude is automatically processed as a

byproduct, as outlined in the Introduction, and with a more gross classification in MC, where magnitude is intentionally processed and classified, although the gross classification seems not to fully overrule the more fine-graded representation. Moreover, the findings speak in favor of Gevers, Verguts, et al.'s (2006) theory that the combination of a Numerical Distance Effect, which in our data received strong evidence (Exploratory 4), and a typically larger SNARC effect for slower responses lead to a more categorical MC-SNARC. Note that the fit of the categorical vs. continuous models did not differ much within each task in absolute terms ($R^2_{continous}$ = 0.42 vs. $R^2_{categorical}$ = 0.47 in MC, and $R^2_{continous}$ = 0.20 vs. $R^2_{categorical}$ = 0.16 in PJ), which might partly be due to the high collinearity (*r* = .913) between the categorical and continuous predictors. Interestingly, the proportion R^2 of explained variance is much larger in MC than in PJ in either model (see fit of the regression lines in Figures 4 vs. 5). In the literature, a large range of reported R^2 can be found (ranging from 0.14 to 0.93 for the continuous and from 0.57 to 0.99 for the categorical MC-SNARC in adult participants, see Table A1 in Appendix A).

A Numerical Size Effect was also detected in both tasks, as expected (Hypothesis 3a). However, while there was strong evidence for it in PJ with the minimal sample size of 500 participants already, moderate evidence for it in MC was only found after increasing the sample size in four steps of 50 participants each, so that this optional recruitment stopping criterion was only met at a sample size of 700. Unexpectedly, the Numerical Size Effect was stronger in PJ than in MC (Hypothesis 3b). We had hypothesized that the effect would be stronger in MC, where processing magnitude was task-relevant, than in PJ, where it was task-irrelevant. A potential explanation for this result is that responses are generally faster in MC than in PJ, and the Numerical Size Effect has been shown to be stronger in slower responses (Hohol et al., 2020; see Supplementary Material, Table S3 there). We therefore checked

separately in each task whether a correlation between the Numerical Size Effect and overall response speed could be found; however, our data revealed strong evidence against such a correlation (Exploratory 13, Appendix B), ruling out this explanation. A different potential explanation is that all number magnitudes should be paid attention to equally strongly when performing MC, attenuating the Numerical Size Effect, whereas number magnitude processing is not intentional but rather automatic in PJ. Importantly, the processing of small number magnitudes is more automatic than the processing of large number magnitudes, thus enhancing the Numerical Size Effect in PJ.

Findings obtained from the exploratory analysis

First, we found no task-order effects (Exploratory 1), at least when averaging across participants, which is a relevant finding for future studies in which both MC and PJ in within-subjects designs will be used. Evidence regarding compatibility-order effects on the PJ were inconclusive (Exploratory 2), being in line with null effects reported in several studies (Bulut et al., 2025; Cipora, van Dijck, et al., 2019; Moorkens et al., 2025; Roth, Jordan, et al., 2024). However, we detected compatibility-order effects showing that the MC-SNARC was stronger when participants completed the incompatible block first and the compatible block second, than when participants did the task in the reversed order (in line with findings by Bulut et al., 2025, but opposed to findings by van Galen & Reitsma, 2008). This is a typical pattern, because when the first block is incompatible, participants are particularly slow for two reasons (see Roth et al., 2024): They have to familiarize themselves with the task (Exploratory 8, Appendix B), and they additionally need to overcome their natural response bias due to the SNARC effect. When the first block is compatible, however, RT are more similar between blocks, because participants familiarize themselves with the task in the first block, whereas they need to overcome their natural response bias due to the SNARC

effect in the second block. Because there were only few dropouts, we did not check for motivational confounding (Reips, 2002; Reips, Overlander & Bannert, 2025), which may have resulted from the compatibility-order asymmetry. For compatibility effects on general RT, please refer to Exploratory 11 and 12 (Appendix B).

As discussed earlier, the shape of the SNARC effect is not necessarily purely categorical or purely continuous (see Panel C in Figure 1), especially in MC. However, Exploratories 3 and 5 (investigating a potential mixed shape overall and separately for early and late trials) were not able to quantify such a mixture adequately due to high collinearity between the categorical and continuous predictors, and due to the existence of only eight data points. In Exploratories 3 and 5, more evidence was found for the continuous predictor in both tasks, which Figures 4 and 5 support. In contrast, the categorical fit was slightly higher than the continuous fit for MC in the results regarding Hypothesis 1, where categorical and continuous magnitude predictors were tested against each other by comparing two separate models. In total, this shows that the shape of the SNARC effect is not clearcut in MC and therefore sensitive to the analysis used. For given reasons, we advise researchers to analyze the MC-SNARC with the following steps: First, they should compute a categorical model (Hypothesis 1), and second, if visual inspection of dRT suggests a continuous proportion of the MC-SNARC, they should additionally test continuous slopes separately for small and large numbers (Exploratory 17).

Regarding the categorization into “small” and “large” numbers, we compared the fit of five regression models using different boundaries for each task (Exploratory 4). We suspected that number 5 would be the boundary in most participants in MC, where it is explicitly defined as such according to the task instructions, while there might not be any clearcut boundary in PJ. We found that number 5 was indeed the most likely underlying boundary for about half of the sample in MC. Adjacent

boundaries (3.5 and 6.5) were descriptively still more prominent than most extreme boundaries (2.5 and 7.5), leading to a unimodal distribution around 5 in MC (see Table 5). In contrast, the distribution of boundaries seemed rather random in the analysis for PJ, so that descriptively, number 5 even happened to be the least favored boundary within the sample. However, not only does the most likely underlying boundary vary considerably between individuals, especially in PJ, but also it seems to be rather unstable within individuals (see Figure 6): The split-half reliability of the boundaries could not be distinguished from 0 in either of the tasks. Note that the stability was rather high for those participants revealing the boundary 5 in MC (see Figure 6). Expectedly, when comparing the continuous model with the categorical model using the favored boundary for each participant, the categorical model still fits the data better than the continuous one in MC (as in Hypothesis 1). More interestingly, this was also the case for PJ, suggesting that the underlying shape of the spatial number representation might be categorical in that task as well, and just vary a lot more between participants. One could therefore think that the better fit of the continuous model found in the confirmatory analysis (Hypothesis 2) might only result from large interindividual differences in categorical boundaries and averaging across various favored models. However, taking into account the null split-half reliability discussed above and often dramatic changes in categorical boundaries in PJ between the halves suggests that the underlying model is in fact continuous.

The split-half reliability of the MC-SNARC was rather high ($r = 0.89$), whereas the split-half reliability of the PJ-SNARC was rather low ($r = 0.27$), and also lower than previously reported in the literature (Exploratory 4, Figure 7). These differences between tasks might be caused by the MC-SNARC resulting from an aggregation of only two possible values for the magnitude predictor, whereas the PJ-SNARC results from eight datapoints.

When quantified with a continuous predictor (although a categorical one fits better in MC) the MC-SNARC and the PJ-SNARC were similarly strong (Exploratory 6) and weakly correlated with one another (Exploratory 7). A stronger MC-SNARC than PJ-SNARC would have been expected in light of stronger magnitude processing in MC than in PJ: In terms of the dual-route model by Gevers, Ratinckx, et al. (2006), both the slow conditional and the fast unconditional route are activated in MC, but only the fast unconditional route is activated in PJ. In other words, one of both prerequisites for the SNARC (as defined by Cipora, He, & Nuerk, 2020) is fulfilled in MC anyways because of the task instructions, namely magnitude processing, so that only the other one needs to happen automatically, namely the directional association. However, the opposite prediction (i.e., a stronger PJ-SNARC than MC-SNARC) could be derived from longer RT in PJ than in MC and a typically stronger SNARC in slower responses. As outlined in the Introduction, both reasonings are plausible, and the observed lack of a difference between the MC-SNARC and the PJ-SNARC might be due to both mechanisms cancelling each other out. The positive correlation between the MC-SNARC and the PJ-SNARC is interesting in light of the literature, where, to our knowledge, a correlation had only been found to be significant by Cheung et al. (2015) and by Cipora (2014; but only in a unimanual setup), while null results were obtained in all other studies (Bulut et al., 2025; Cipora, 2014, and Cipora et al., 2016, for PJ, and Hohol et al., 2020, for MC; Didino et al., 2019; Fattorini et al., 2015, Study 1; Georges et al., 2017; van Dijck and Doricchi; 2019). For an overview, please refer to the rightmost column in Table A1 (Appendix A). Arguments against a correlation might be low intraindividual stability, low reliability, or low internal consistency of the SNARC effect in both tasks. However, the results of the present study might be more trustworthy thanks to the large sample size, and previous studies might just have failed to detect an existing but small correlation. The observed correlation leads to the

conclusion that the mental representation of numbers underlying in MC and PJ overlaps at least partly.

In addition to what was proposed in the Stage-1 Registered Report, the large amount of data permitted us to investigate several other aspects of MC and PJ (see Appendix B). When habituating to such tasks, participants usually get faster (i.e., training effects). This was reflected by faster responses in the second than in the first block for the case of MC (Exploratory 8, Appendix B), and by faster responses in the second than first half of each block for the case of PJ (Exploratory 9, Appendix B).

A typical Odd effect (i.e., faster responses to even than to odd numbers) was only observed in PJ (Exploratory 10, Appendix B). A reversed Odd effect was observed in MC, which seems striking at first sight. However, this is fully in line with the large Numerical Distance Effect observed in MC (Replication Check 4): The numbers closest to the reference number 5 (i.e., 4 and 6) are responded to more slowly than the numbers with the greatest numerical distance to the reference number 5 (i.e., 1 and 9); and crucially, the first happen to be even, whereas the latter happen to be odd. Thus, it seems that the Numerical Distance Effect is overruling the Odd effect, which might not be as strong in MC as in PJ anyways (see Figure 3).

A typical effect on overall RT was observed regarding the SNARC-compatibility in each task, such that responses were faster for SNARC-compatible than for SNARC-incompatible trials (by 10 ms in MC and by 8 ms in PJ; Exploratory 11, Appendix B). No such MARC-compatibility effect on overall RT could be found, with RT being similar for MARC-compatible and MARC-incompatible trials in each task (Exploratory 12, Appendix B). This null result (in Bayesian sense) is in line with the lack of a MARC effect in both tasks (Replication Check 2), and also with the lack of a compatibility-order effect on the PJ-MARC, where MARC-compatible and MARC-incompatible trials

were presented in two separate blocks (Exploratory 2). Note that the null MARC effect did not differ between left- and right-handers (Exploratory 16, Appendix B).

Lastly, Moretti et al. (2025) recently suggested to focus on trials that elicit the highest response conflict when measuring cognitive effects. We therefore re-analyzed the PJ-SNARC and re-checked its split-half reliability while only considering post-SNARC-compatible trials. The effect size of the continuous PJ-SNARC was much larger when excluding post-SNARC-incompatible trials (Exploratory 14, Appendix B) than when keeping them (Hypothesis 2); however, its split-half reliability remained similar (Exploratory 15, Appendix B).

Summary and conclusion

In this large-scale online study, the SNARC effect was replicated in the two most commonly used tasks (namely, MC and PJ). Analyzing data from over 600 participants, a rather categorical shape of the MC-SNARC with some continuous proportion and a rather continuous shape of the PJ-SNARC were found. Assessing these two most widely used SNARC tasks in a within-subjects design also permitted to detect a small but positive correlation between the MC-SNARC and the PJ-SNARC. Moreover, we thoroughly investigated numerous other numerical-cognition and general-RT phenomena, laying a comprehensive groundwork for further investigations using MC and PJ.

The most important practical implication of this work, given that statistical models should align with scientific ones (Westermann & Hager, 2017), is that researchers should use a categorical statistical model to quantify the SNARC effect in MC. They might even consider computing different categorical models with individual boundaries between “small” and “large” numbers, and additionally investigate continuous aspects. At the same time, researchers can continue using the traditional continuous model in PJ. The most important theoretical implication of this work is that,

instead of speaking about *the* SNARC effect, researchers should rather acknowledge that different SNARC effects exist, and that the shape and other properties of the phenomenon depend on the used task.

Data and code availability

All R scripts for data preprocessing and analysis as well as all anonymized datasets are made publicly available via the Open Science Framework (https://osf.io/g48s2/).

Author contributions

All the authors have full access to all the data and take responsibility for the integrity of the data and the accuracy of the data analysis. *Conceptualization*: K. Cipora, H.-C. Nuerk, U.-D. Reips; *Data Curation*: K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips, L. Roth; *Formal Analysis*: K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips, L. Roth; *Funding Acquisition*: K. Cipora, H.-C. Nuerk, U.-D. Reips.; *Investigation*: K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips, L. Roth; *Methodology*: K. Cipora, H.-C. Nuerk, U.-D. Reips, L. Roth; *Project Administration*: H.-C. Nuerk, U.-D. Reips, L. Roth; *Resources*: H.-C. Nuerk, U.-D. Reips; *Software*: A. T. Overlander, U.-D. Reips; *Supervision*: K. Cipora, H.-C. Nuerk, U.-D. Reips; *Validation*: K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips, L. Roth; *Visualization*: L. Roth; *Writing - original draft*: L. Roth; *Writing - review and editing:* K. Cipora, H.-C. Nuerk, A. T. Overlander, U.-D. Reips.

Competing interests

The authors declare no conflicts of interest with the content of this article.

Acknowledgements

This research was supported by the DFG project “Replicability of Fundamental Results on Spatial-Numerical Associations in Highly Powered Online Experiments (e-SNARC)” (NU 265/8-1 and RE 2655/3-1) granted to Hans-Christoph Nuerk and Ulf-Dietrich Reips, supporting Lilly Roth and Annika Tave Overlander, with the assistance of Krzysztof Cipora as a cooperation partner. Hans-Christoph Nuerk’s work on spatial-numerical associations is additionally supported by the DFG-projects 265-5/1 and 265-5/2: On the interplay of modal and amodal encodings underlying space-metric associations (SMAs). Krzysztof Cipora is supported by the UKRI Economic and Social Research Council (grant number ES/W002914/1). The authors would like to thank Sebastian Sandbrink for English language proofreading of the Stage-1 Registered Report.

References

Abrahamse, E., van Dijck, J.-P., & Fias, W. (2016). How does working memory enable number-induced spatial biases? *Frontiers in Psychology, 7*, Article 977. https://doi.org/10.3389/fpsyg.2016.00977

Bachot, J., Gevers, W., Fias, W., & Roeyers, H. (2005). Number sense in children with visuospatial disabilities: Orientation of the mental number line. *Psychology Science*, *47*(1), 172–183.

Bae, G. Y., Choi, J. M., Cho, Y. S., & Proctor, R. W. (2009). Transfer of magnitude and spatial mappings to the SNARC effect for parity judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1506–1521. https://doi.org/10.1037/a0017257

Banks, W. P., Fujii, M., & Kayra-Stuart, F. (1976). Semantic congruity effects in comparative judgments of magnitudes of digits. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(3), 435–447. https://doi.org/10.1037/0096-1523.2.3.435

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Hocking, T., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X., Saporta, R., Seiskari, O., Dong, X., Lang, M., … Czekanski, M. (2024). *data.table: Extension of „data.frame“* [R package, version 1.17.0]. https://cran.r-project.org/web/packages/data.table/index.html

Basso Moro, S., Dell'Acqua, R., & Cutini, S. (2018). The SNARC effect is not a unitary phenomenon. *Psychonomic Bulletin & Review*, *25*(2), 688–695. https://doi.org/10.3758/s13423-017-1408-3

Brunson, J.C., & Read, Q.D. (2023). *ggalluvial: Alluvial Plots in 'ggplot2'* [R package, version 0.12.5]. https://CRAN.R-project.org/package=ggalluvial

Brysbaert, M. (1995). Arabic number reading: On the nature of the numerical scale and the origin of phonological recoding. *Journal of Experimental Psychology: General*, *124*(4), 434–452. https://doi.org/10.1037/0096-3445.124.4.434

Buerkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software.* *80*(1), 1-28. https://doi.org/10.18637/jss.v080.i01

Bull, R., Marschark, M., & Blatto-Vallee, G. (2005). SNARC hunting: Examining number representation in deaf students. *Learning and Individual Differences*, *15*(3), 223–236. https://doi.org/10.1016/j.lindif.2005.01.004

Bulut, M., Roth, L., Bahreini, N., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2025). One direction? Cultural aspects of the mental number line beyond reading direction. *Psychological Research, 89,* Article 37. https://doi.org/10.1007/s00426-024-02038-4

Casasanto, D., & Pitt, B. (2019). The faulty magnitude detector: Why SNARC-like tasks cannot support a generalized magnitude system. *Cognitive Science, 43*(10), Article e12794. https://doi.org/10.1111/cogs.12794

Cheung, C.-N., Ayzenberg, V., Diamond, R. F. L., Yousif, S., & Lourenco, S. F. (2015). Probing the mental number line: A between-task analysis of spatial-numerical associations. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 357–362.

Cipora, K. (2014). *Between-task consistency of the SNARC effect* [poster presentation]. XXXIInd European Workshop on Cognitive Neuropsychology, Bressanone, Italy. https://osf.io/nx6vq/

Cipora, K., & Göbel, S. M. (2013). *Number-space associations: Just how reliable is the SNARC effect* [poster presentation]. In 16th European Workshop on Cognitive Neuropsychology. https://osf.io/nx6vq

Cipora, K., Haman, M., Domahs, F., & Nuerk, H.-C. (2020). Editorial: On the development of space-number relations: Linguistic and cognitive determinants, influences, and associations. *Frontiers in Psychology*, *11*, Article 182. https://doi.org/10.3389/fpsyg.2020.00182

Cipora, K., Patro, K., & Nuerk, H.-C. (2018). Situated influences on spatial–numerical associations. In T. L. Hubbard (Ed.), *Spatial biases in perception and cognition* (pp. 41–59). Cambridge University Press. https://doi.org/10.1017/9781316651247.004

Cipora, K., Soltanlou, M., Reips, U.-D., & Nuerk, H.-C. (2019). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behavior Research Methods*, *51*(4), 1676–1692. https://doi.org/10.3758/s13428-019-01213-5

Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Goebel, S. M., Willmes, K., Pesenti, M., Schiltz, C., & Nuerk, H.-C. (2019). *A minority pulls the sample mean: on the individual prevalence of robust group-level cognitive phenomena – the instance of the SNARC effect* [preprint]. Open Science Framework. https://doi.org/10.31234/osf.io/bwyr3

Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Schiltz, C., Pesenti, M., Goebel, S. M., Willmes, K., & Nuerk, H.-C. (2020). *Parity and space: On the prevalence and individual differences in the MARC effect [poster presentation]*. XXXVIIIth European Workshop on Cognitive Neuropsychology, Bressanone, Italy. https://osf.io/4qhpw

Cipora, K., & Wood, G. (2017). Finding the SNARC instead of hunting it: A 20∗ 20 Monte Carlo investigation. *Frontiers in Psychology, 8*, 243273. https://doi.org/10.3389/fpsyg.2017.01194

Cohen, J. (1988). The effect size. In *Statistical power analysis for the behavioral sciences* (2nd ed., p. 77–83). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Sciences, 24*(3), 200–207. https://doi.org/10.1016/j.tics.2019.12.009

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371–396. https://doi.org/10.1037/0096-3445.122.3.371

Deng, Z., Chen, Y., Zhang, M., Li, Y., & Zhu, X. (2018). The association of number and space under different tasks: Insight from a process perspective. *Frontiers in Psychology*, *9*, Article 957. https://doi.org/10.3389/fpsyg.2018.00957

Deng, Z., Chen, Y., Zhu, X., & Li, Y. (2017). The effect of working memory load on the SNARC effect: Maybe tasks have a word to say. *Memory & Cognition*, *45*(3), 428–441. https://doi.org/10.3758/s13421-016-0676-x

Didino, D., Breil, C., & Knops, A. (2019). The influence of semantic processing and response latency on the SNARC effect. *Acta Psychologica*, *196*, 75–86. https://doi.org/10.1016/j.actpsy.2019.04.008

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice*, *8*(1), 9–26. https://doi.org/10.1037/cns0000258

Fattorini, E., Pinto, M., Rotondaro, F., & Doricchi, F. (2015). Perceiving numbers does not cause automatic shifts of spatial attention. *Cortex*, *73*, 298–316. https://doi.org/10.1016/j.cortex.2015.09.007

Fias, W., Brysbaert, M., Geypens, F., & D'Ydewalle, G. (1996). The importance of magnitude information in numerical processing: Evidence from the SNARC

effect. *Mathematical Cognition*, *2*(1), 95–110. https://doi.org/10.1080/135467996387552

Fias, W., Lammertyn, J., Reynvoet, B., Dupont, P., & Orban, G. A. (2003). Parietal representation of symbolic and nonsymbolic magnitude. *Journal of Cognitive Neuroscience*, *15*(1), 47–56. https://doi.org/10.1162/089892903321107819

Fischer, M. H., & Shaki, S. (2014). Spatial associations in numerical cognition: From single digits to arithmetic. *Quarterly Journal of Experimental Psychology*, *67*(8), 1461–1483. https://doi.org/10.1080/17470218.2014.927515

Fitousi, D., Shaki, S., & Algom, D. (2009). The role of parity, physical size, and magnitude in numerical cognition: The SNARC effect revisited. *Attention, Perception & Psychophysics*, *71*(1), 143–155. https://doi.org/10.3758/APP.71.1.143

Garaizar, P., & Reips, U.-D. (2019). Best practices: Two web browser-based methods for stimulus presentation in behavioral experiments with high resolution timing requirements. *Behavior Research Methods, 51*, 1441–1453. https://doi.org/10.3758/s13428-018-1126-4

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Georges, C., Hoffmann, D., & Schiltz, C. (2013). *The SNARC effect and its relationship to spatial abilities in women* [poster presentation]. In 18th conference of the European Society for Cognitive Psychology, Budapest, Hungary. http://hdl.handle.net/10993/13012

Georges, C., Hoffmann, D., & Schiltz, C. (2017). How and why do number-space associations co-vary in implicit and explicit magnitude processing tasks?

Journal of Numerical Cognition, *3*(2), 182–211. https://doi.org/10.5964/jnc.v3i2.46

Gevers, W., Ratinckx, E., Baene, W. de, & Fias, W. (2006). Further evidence that the SNARC effect is processed along a dual-route architecture: Evidence from the lateralized readiness potential. *Experimental Psychology*, *53*(1), 58-68. https://doi.org/10.1027/1618-3169.53.1.58

Gevers, W., Santens, S., Dhooge, E., Chen, Q., van den Bossche, L., Fias, W., & Verguts, T. (2010). Verbal-spatial and visuospatial coding of number-space interactions. *Journal of Experimental Psychology: General, 139*(1), 180–190. https://doi.org/10.1037/a0017688

Gevers, W., Verguts, T., Reynvoet, B., Caessens, B., & Fias, W. (2006). Numbers and space: A computational model of the SNARC effect. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 32–44. https://doi.org/10.1037/0096-1523.32.1.32

Gibson, L. C., & Maurer, D. (2016). Development of SNARC and distance effects and their relation to mathematical and visuospatial abilities. *Journal of Experimental Child Psychology*, *150*, 301–313. https://doi.org/10.1016/j.jecp.2016.05.009

Gökaydin, D., Brugger, P., & Loetscher, T. (2018). Sequential Effects in SNARC. *Scientific Reports*, *8*(1), Article 10996. https://doi.org/10.1038/s41598-018-29337-2

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General*, *121*(4), 480–506. https://doi.org/10.1037/0096-3445.121.4.480

Han, M., Mao, X., Cai, M., Jia, X., & Guo, C. (2017). The effect of positive and negative signs on the SNARC effect in the magnitude judgment task. *Acta Psychologica Sinica*, *49*(8), 995–1008. https://doi.org/10.3724/SP.J.1041.2017.00995

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Herrera, A., Macizo, P., & Semenza, C. (2008). The role of working memory in the association between number magnitude and space. *Acta Psychologica*, *128*(2), 225–237. https://doi.org/10.1016/j.actpsy.2008.01.002

Hines, T. M. (1990). An odd effect: Lengthened reaction times for judgments about odd digits. *Memory & Cognition*, *18*(1), 40–46. https://doi.org/10.3758/BF03202644

Hoffmann, D., Hornung, C., Martin, R., & Schiltz, C. (2013). Developing number-space associations: SNARC effects using a color discrimination task in 5-year-olds. *Journal of Experimental Child Psychology*, *116*(4), 775–791. https://doi.org/10.1016/j.jecp.2013.07.013

Hohol, M., Willmes, K., Nęcka, E., Brożek, B., Nuerk, H.-C., & Cipora, K. (2020). Professional mathematicians do not differ from others in the symbolic numerical distance and size effects. *Scientific Reports*, *10*(1), Article 11531. https://doi.org/10.1038/s41598-020-68202-z

Hubbard, E. M., Ranzini, M., Piazza, M., & Dehaene, S. (2009). What information is critical to elicit interference in number-form synaesthesia? *Cortex*, *45*(10), 1200–1216. https://doi.org/10.1016/j.cortex.2009.06.011

Ito, Y., & Hatta, T. (2004). Spatial structure of quantitative representation of numbers: Evidence from the SNARC effect. *Memory & Cognition*, *32*(4), 662–673. https://doi.org/10.3758/BF03195857

Kassambara, A. (2025). *ggpubr: „ggplot2“ based publication ready plots [R package, version 0.6.1]*. https://CRAN.R-project.org/package=ggpubr

Kelter, R. (2021). Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Computational Statistics*, *36*(2), 1263–1288. https://doi.org/10.1007/s00180-020-01034-7

Kiesel, A., Wendt, M., & Peters, A. (2007). Task switching: On the origin of response congruency effects. *Psychological Research*, *71*(2), 117–125. https://doi.org/10.1007/s00426-005-0004-8

Koch, N. N., Huber, J. F., Lohmann, J., Cipora, K., Butz, M. V., & Nuerk, H.-C. (2023). Mental number representations are spatially mapped both by their magnitudes and ordinal positions. *Collabra: Psychology*, *9*(1), Article 67908. https://doi.org/10.1525/collabra.67908

Lohmann, J., Schroeder, P. A., Nuerk, H.-C., Plewnia, C., & Butz, M. V. (2018). How deep is your SNARC? Interactions between numerical magnitude, response hands, and reachability in peripersonal space. *Frontiers in Psychology*, *9*, Article 622. https://doi.org/10.3389/fpsyg.2018.00622

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *16*(1), 149–157. https://doi.org/10.1037/0278-7393.16.1.149

Microsoft Corporation. (2024). PowerPoint for Microsoft 365 (Version 2408) [Computer software]. Microsoft. https://www.microsoft.com

Moorkens, J., Dijck, J.-P. van, & Fias, W. (2025). The parity judgment SNARC effect: The role of response mapping order and the nature of the instruction. *Journal of Numerical Cognition, 11*, 1–12. https://doi.org/10.5964/jnc.15051

Moretti, L., Koch, I., Hornjak, R., & von Bastian, C. C. (2025). Quality over quantity: Focusing on high-conflict trials to improve the reliability and validity of attentional control measures. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* https://doi.org/10.1037/xlm0001466

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2024). *BayesFactor: Computation of Bayes factors for common designs [R package, version 0.9.12-4.7]*. https://CRAN.R-project.org/package=BayesFactor

Mourad, A., & Leth-Steensen, C. (2017). Spatial reference frames and SNARC. *Journal of Cognitive Psychology*, *29*(2), 113–128. https://doi.org/10.1080/20445911.2016.1249483

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–1520. https://doi.org/10.1038/2151519a0

Nathan, M. B., Shaki, S., Salti, M., & Algom, D. (2009). Numbers and space: Associations and dissociations. *Psychonomic Bulletin & Review*, *16*(3), 578–582. https://doi.org/10.3758/PBR.16.3.578

Nuerk, H.-C., Bauer, F., Krummenacher, J., Heller, D., & Willmes, K. (2005). The power of the mental number line: How the magnitude of unattended numbers affects performance in an Eriksen task. *Psychology Science*, *47*(1), 34–50. https://psycnet.apa.org/record/2005-11470-005

Nuerk, H.-C., Iversen, W., & Willmes, K. (2004). Notational modulation of the SNARC and the MARC (linguistic markedness of response codes) effect. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *57*(5), 835–863. https://doi.org/10.1080/02724980343000512

Nuerk, H.-C., & Schroeder, P. A. (2024). *A Psychological Markedness Account of human cognition: The example of numerical cognition* [manuscript in preparation]. Department of Psychology, University of Tuebingen, Germany.

Nuerk, H.-C., Wood, G., & Willmes, K. (2005). The universal SNARC effect: The association between number magnitude and space is amodal. *Experimental Psychology*, *52*(3), 187–194. https://doi.org/10.1027/1618-3169.52.3.187

Patro, K., Nuerk, H. C., Cress, U., & Haman, M. (2014). How number-space relationships are assessed before formal schooling: A taxonomy proposal. *Frontiers in psychology, 5*, Article 55976. https://doi.org/10.3389/fpsyg.2014.00419

Pfister, R., Schroeder, P. A., & Kunde, W. (2013). SNARC struggles: Instant control over spatial-numerical associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1953–1958. https://doi.org/10.1037/a0032991

Pinto, M., Pellegrino, M., Lasaponara, S., Scozia, G., D'Onofrio, M., Raffa, G., Nigro, S., Arnaud, C. R., Tomaiuolo, F., & Doricchi, F. (2021). Number space is made by response space: Evidence from left spatial neglect. *Neuropsychologia*, *154*, Article 107773. https://doi.org/10.1016/j.neuropsychologia.2021.107773

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, *132*(3), 416–442. https://doi.org/10.1037/0033-2909.132.3.416

Prpic, V., Fumarola, A., Tommaso, M. de, Luccio, R., Murgia, M., & Agostini, T. (2016). Separate mechanisms for magnitude and order processing in the spatial-numerical association of response codes (SNARC) effect: The strange case of musical note values. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1241–1251. https://doi.org/10.1037/xhp0000217

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.4.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reips, U.-D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological Experiments on the Internet* (pp. 89–117). Academic Press. https://doi.org/10.1016/B978-012099980-4/50005-8

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. https://doi.org/10.1026//1618-3169.49.4.243

Reips, U.-D. (2007). The methodology of Internet-based experiments. In A. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 373–390). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199561803.013.0024

Reips, U.-D. (2009). Internet experiments: Methods, guidelines, metadata. In B. E. Rogowitz & T. N. Pappas (Eds.), *SPIE Proceedings, Human Vision and Electronic Imaging XIV,* 724008. SPIE. https://doi.org/10.1117/12.823416

Reips, U.-D. (2021). Web-based research in psychology: A review. *Zeitschrift für Psychologie, 229*(4), 198–213. https://doi.org/10.1027/2151-2604/a000475

Reips, U.-D., & Neuhaus, C. (2002). WEXTOR: A Web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, & Computers*, *34*(2), 234–240. https://doi.org/10.3758/bf03195449

Reips, U.-D., Overlander, A. T., & Bannert, M. (2025). Dropout analysis: A method for data from Internet-based research and dropR, an R-based web app and package to analyze and visualize dropout. *Behavior Research Methods, 57*(1), Article 231. https://doi.org/10.3758/s13428-025-02730-2

Roettger, T. B., & Domahs, F. (2015). Grammatical number elicits SNARC and MARC effects as a function of task demands. *Quarterly Journal of Experimental Psychology (2006)*, *68*(6), 1231–1248. https://doi.org/10.1080/17470218.2014.979843

Roth, L., Caffier, J., Reips, U.-D., Cipora, K., Braun, L., & Nuerk, H.-C. (2025a). True colors SNARC: Semantic number processing is highly automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001431

Roth, L., Caffier, J., Reips, U.-D., Nuerk, H.-C., Overlander, A. T., & Cipora, K. (2025b). One and only SNARC? Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude. *Royal Society Open Science, 12*(1), Article 241585. https://doi.org/10.1098/rsos.241585

Roth, L., Huber, J., Kronenthaler, S., van Dijck, J., Cipora, K., Butz, M. V., & Nuerk, H.-C. (2025c). Looks like SNARC spirit: Coexistence of short- and long-term associations between letters and space. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218251324437

Roth, L., Jordan, V., Schwarz, S., Willmes, K., Nuerk, H.-C., van Dijck, J.-P., & Cipora, K. (2024). Don't SNARC me now! Intraindividual variability of cognitive phenomena – Insights from the Ironman paradigm. *Cognition, 248*, Article 105781. https://doi.org/10.1016/j.cognition.2024.105781

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989X.13.1.19

Saeki, E., & Saito, S. (2009). Verbal representation in task order control: An examination with transition and task cues in random task switching. *Memory & Cognition*, *37*(7), 1040–1050. https://doi.org/10.3758/MC.37.7.1040

Santens, S., & Gevers, W. (2008). The SNARC effect does not imply a mental number line. *Cognition*, *108*(1), 263–270. https://doi.org/10.1016/j.cognition.2008.01.002

Schiller, F., Eloka, O., & Franz, V. H. (2016). Using key distance to clarify a theory on the SNARC. *Perception*, *45*(1-2), 196–221. https://doi.org/10.1177/0301006615616754

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017). Switching between multiple codes of SNARC-like associations: Two conceptual replication attempts with anodal tDCS in sham-controlled cross-over design. *Frontiers in Neuroscience, 11*, Article 654. https://doi.org/10.3389/fnins.2017.00654

Schwarz, W., & Keus, I. M. (2004). Moving the eyes along the mental number line: Comparing SNARC effects with saccadic and manual responses. *Perception & Psychophysics*, *66*(4), 651–664. https://doi.org/10.3758/bf03194909

Schwarz, W., & Müller, D. (2006). Spatial associations in number-related tasks: A comparison of manual and pedal responses. *Experimental Psychology*, *53*(1), 4–15. https://doi.org/10.1027/1618-3169.53.1.4

Shaki, S., & Gevers, W. (2011). Cultural characteristics dissociate magnitude and ordinal information processing. *Journal of Cross-Cultural Psychology*, *42*(4), 639–650. https://doi.org/10.1177/0022022111406100

Tlauka, M. (2002). The processing of numbers in choice-reaction tasks. *Australian Journal of Psychology*, *54*(2), 94–98. https://doi.org/10.1080/00049530210001706553

Toomarian, E. Y., & Hubbard, E. M. (2018). On the genesis of Spatial-Numerical Associations: Evolutionary and cultural factors co-construct the Mental Number Line. *Neuroscience & Biobehavioral Reviews*, *90*, 184–199. https://doi.org/10.1016/j.neubiorev.2018.04.010

Tzelgov, J. (1997). Specifying the relations between automaticity and consciousness: A theoretical note. *Consciousness and Cognition*, *6*(2-3), 441–451. https://doi.org/10.1006/ccog.1997.0303

Tzelgov, J., Ganor-Stern, D., Kallai, A. Y., & Pinhas, M. (2015). Primitives and non-primitives of numerical representations. In R. C. Kadosh & A. Dowker (Eds.), *Oxford library of psychology: The Oxford handbook of numerical cognition* (pp. 45–66). Oxford University Press.

Tzelgov, J., Meyer, J., & Henik, A. (1992). Automatic and intentional processing of numerical information. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *18*(1), 166–179. https://doi.org/10.1037/0278-7393.18.1.166

Van Dijck, J.-P., Abrahamse, E. L., Acar, F., Ketels, B., & Fias, W. (2014). A working memory account of the interaction between numbers and spatial attention. *Quarterly Journal of Experimental Psychology*, *67*(8), 1500–1513. https://doi.org/10.1080/17470218.2014.903984

Van Dijck, J.-P., & Doricchi, F. (2019). Multiple left-to-right spatial representations of number magnitudes? Evidence from left spatial neglect. *Experimental Brain Research*, *237*(4), 1031–1043. https://doi.org/10.1007/s00221-019-05483-5

Van Dijck, J.-P., Gevers, W., & Fias, W. (2009). Numbers are associated with different types of spatial information depending on the task. *Cognition*, *113*(2), 248–253. https://doi.org/10.1016/j.cognition.2009.08.005

Van Dijck, J.-P., Gevers, W., Lafosse, C., & Fias, W. (2012). The heterogeneous nature of number-space interactions. *Frontiers in Human Neuroscience*, *5*, Article 182. https://doi.org/10.3389/fnhum.2011.00182

Van Galen, M. S., & Reitsma, P. (2008). Developing access to number magnitude: A study of the SNARC effect in 7- to 9-year-olds. *Journal of Experimental Child Psychology*, *101*(2), 99–113. https://doi.org/10.1016/j.jecp.2008.05.001

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Viarouge, A., Hubbard, E. M., & McCandliss, B. D. (2014). The cognitive mechanisms of the SNARC effect: An individual differences approach. *PLOS ONE*, *9*(4), Article e95756. https://doi.org/10.1371/journal.pone.0095756

Weis, T., Nuerk, H.-C., & Lachmann, T. (2018). Attention allows the SNARC effect to operate on multiple number lines. *Scientific Reports*, *8*(1), Article 13778. https://doi.org/10.1038/s41598-018-32174-y

Westermann, R., & Hager, W. (1983). On severe tests of trend hypotheses in psychology. *The Psychological Record*, *33*(2), 201–211. https://doi.org/10.1007/BF03394838

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., van den Brand, T., & Posit PBC (2024). *ggplot2: Create elegant data visualisations using the grammar of graphics [R package, version 3.5.2].* https://CRAN.R-project.org/package=ggplot2

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation [R package, version 1.1.4].* https://cran.r-project.org/web/packages/dplyr/index.html

Wickham, H., Pedersen, T. L., Seidel, D., & Posit Software. (2025). *scales: Scale Functions for Visualization [R package, version 1.4.0]*. https://cran.r-project.org/web/packages/scales/index.html

Wickham, H., Vaughan, D., Girlich, M., & Ushey, K. (2023). *tidyr: Tidy messy data [R package, version 1.3.1]*. https://cran.r-project.org/web/packages/tidyr/index.html

Wood, G., Willmes, K., Nuerk, H.-C., & Fischer, M. H. (2008). On the cognitive link between space and number: A meta-analysis of the SNARC effect. *Psychology Science*, *50*(4), 489–525.

Xiang, X., Yan, L., Fu, S., & Nan, W. (2022). *Processing stage flexibility of the SNARC effect: Task relevance or magnitude relevance?* [preprint]. BioRxiv. https://doi.org/10.1101/2022.03.07.482213

Zhang, P., Cao, B., & Li, F. (2022). The role of cognitive control in the SNARC effect: A review. *PsyCh Journal*, *11*(6), 792–803. https://doi.org/10.1002/pchj.586

Zorzi, M., Bonato, M., Treccani, B., Scalambrin, G., Marenzi, R., & Priftis, K. (2012). Neglect impairs explicit processing of the mental number line. *Frontiers in Human Neuroscience*, *6*, Article 125. https://doi.org/10.3389/fnhum.2012.00125

Appendix A

Table A1

Overview of studies including magnitude classification tasks (MC) of visually presented numbers

Study	**Sample**	**Stimuli (and untypical instructions)**	**Repeated-measures regression**		**Correlation between continuous predictors of MC and PJ**
			Continuous predictor	**Categorical predictor**	
Bachot et al. (2005): Control group	16 visuospatially non-impaired children between 7 and 12 years	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -13.97 (*SD* = 26.35), *t(15)* = -2.12, $p < .05$ (one-sided)	Not reported	No PJ
Bae et al. (2009): Practice Task in Experiment 1	28 students	2, 3, 4, 5, 6, 7, 8, 9 (classify small from 2 to 5 or large from 6 to 9)	*slope* = -31.675, $t(7) = -4.40$, $p < .01$, $R^2 = 0.764$	Not reported	No correlation reported
Bull et al. (2005): Experiment 1	20 deaf and 20 hearing students	1, 2, 3, 4, 6, 7, 8, 9 (responses with mouse keys instead of keyboard keys)	Deaf: *slope* = -17.268, $t(7) = -3.21$, $p < .05$, $R^2 = 0.63$ Hearing: *slope* = -21.136, $t(7) = -4.17$, $p < .01$, $R^2 = 0.74$	Not reported	No PJ

Bulut et al. (2025)	130 German, 112 Turkish, and 75 Iranian adults	1, 2, 3, 4, 6, 7, 8, 9	Not reported	German: *slope* = -32.58 (*SD* = 63.38), *t*(129) = -5.86, *p* < .001 Turkish: *slope* = -27.26 (*SD* = 73.84), *t*(111) = -3.91, *p* < .001 Iranian: *slope* = -19.42 (*SD* = 104.92), *t*(74) = -1.60, *p* = .113	German: *r* = -0.02, *p* = .822 Turkish: *r* = -0.08, *p* = .386 Iranian: *r* = 0.10, *p* = .402
Cheung et al. (2015)	125 students	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -5.89 (*SD* = 11.06), *t*(115)= -5.73, *p* < .001	Not reported	*r* = 0.25, *p* < .05
Cipora (2014) and Cipora et al. (2016) for PJ; Hohol et al. (2020, Supplementary Material 3) for MC	100 adults (99 for PJ, and 98 for MC)	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -16.04 (*SD* = 50.57), *t*(97) = -3.14, *p* = .001 (one-sided; calculated from raw data shared by Hohol et al, 2020)	*slope* = -16.0 (*SD* = 50.6), *t*(97) = -3.14, *p* = .001 (one-sided; reported by Hohol et al, 2020)	*r* = 0.06, *p* = .577, 95% confidence interval = [-0.14; 0.25] (PJ-SNARC calculated from raw data shared by Cipora et al., 2019)
Deng et al. (2017): basic 0-load task	112 adults (Exp. 1: 30 spatial + 31 verbal; Exp. 2: 25 spatial + 26 verbal)	1, 2, 3, 4, 6, 7, 8, 9	basic 0-load task without spatial or verbal working memory load: Exp. 1 spatial: *slope* = -6.08 (*SD* = 7.64), *t*(29) = -4.36, *p* < .01	Not reported	No correlation reported

			Exp. 1 verbal: *slope* = -7.09 (*SD* = 10.04), *t*(30) = -3.93, *p* < .01 Exp. 2 spatial: *slope* = -7.59 (*SD* = 16.67), *t*(24) = -2.28, *p* < .05 Exp. 2 verbal: *slope* = -7.33 (*SD* = 17.43), *t*(25) = -2.14, *p* < .05		
Didino et al. (2019)	32 adults	1, 2, 3, 4, 6, 7, 8, 9	Not retained in the stepwise linear regression analysis with forward and backward predictor selection based on the AIC	*slope* = -23.59 (*SE* = 7.39), *t* = -3.19, *p* = 0.019, R^2 = 0.57	*r* = 0.06, 95% confidence interval = [-0.30; 0.40]
Fattorini et al. (2015): Study 1, Supplementary material	60 students	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -7.5 (*SD* = 11.3), *t*(59) = -5.1, *p* < .001	Not reported	*r* = 0.18, *p* = .18, 95% confidence interval = [-0.07; 0.42]
Fitousi et al. (2009): in Experiment 5	16 adults	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -11.652, *F*(1,6) = 20.673, *p* < .01, R^2 = .775	*slope* = -68.951, *F*(1,6) = 52.927, *p* < .01, R^2 = 0.904	No correlation reported
Georges et al. (2017)	90 students	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -5.2 (*SD* = 13.1), *t*(80) = -3.57, *p* = .001	*slope* = -29.18, *t*(1, 6) = -8.33, *p* < .001, R^2 = 0.92	*r* = 0.2, *p* = .07

Gevers et al. (2005): Experiment 2	20 students	1, 2, 3, 4, 6, 7, 8, 9 (presented on the left/right side of the screen but not centrally)	No repeated-measures regression, but instead main effect of SNARC compatibility (compatible: small-left, large-right, incompatible: small-right, large-left): $F(1, 19) = 5.97$, $p < .05$		No correlation reported
Gevers et al. (2006): Experiment 1	40 adults	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -3.52	*slope* = -17.64	No correlation reported
			Stepwise multiple regression reveals a better fit of the categorical predictor: $Z = 2.31$, $p < .05$		
Han et al. (2017): positive numbers	16 students	+1, +2, +3, +4, +6, +7, +8, +9 (numbers presented with plus sign and compared to reference number +5)	Block with positive numbers: *slope* = -14.53, $t(28) = -3.51$, $p < .05$	Not reported	No PJ
Herrera et al. (2008): basic task without working memory load	18 students	1, 2, 3, 4, 6, 7, 8, 9	Basic tasks without phonological or visuospatial working memory load: Exp. 1: *slope* = -7, $t(17) = -2.67$, $p < .05$ Exp. 2: *slope* = -6, $t(17) = -3.33$, $p < .01$	Not reported	No PJ
Hoffmann et al. (2013)	84 children	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -5.50, $t(69) = -0.50$, $p > .30$ (one-sided)	Not reported	No PJ
Hubbard et al. (2009): Experiment 5b	8 adults (non-synaesthetic controls)	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -8.56, $F(1,6) = 19.97$, $p < .005$, $R^2 = 0.77$	Not reported	No correlation reported

Ito and Hatta (2004): Experiment 3	28 students	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -0.08 (*SD* = 6.0), *t*(27) = -0.07, *p* > .4 (one-sided)	Not reported	MC and PJ in between-subjects design
Lohmann et al. (2018)	16 students	1, 2, 3, 4, 6, 7, 8, 9 (responses were given by clinching the left/right fist in virtual reality environment, while manipulating spatial displacement)	Close condition: *slope* = -19.91 (*SD* = 17.7), *t*(24) = 1.50, *p* = .074 (one-sided), R^2 = 0.270 Border condition: *slope* = -18.09 (*SD* = 19.6), *t*(24) = 1.72, *p* = .049 (one-sided), R^2 = 0.282 Hand condition: *slope* = -27.81 (*SD* = 28.9), *t*(24) = 2.36, *p* = .014 (one-sided), R^2 = 0.472 Extrapersonal condition: *slope* = -25.03 (*SD* = 23.9), *t*(24) = 2.05, *p* = .026 (one-sided), R^2 = 0.403	Not reported	No PJ
Mourad and Leth-Steensen (2017): HH and VH conditions	60 students	1, 2, 3, 4, 6, 7, 8, 9 (participants had to imagine a number line from left to right in the HH condition and from bottom to top in the VH condition before MC)	HH: *slope* = -17.46 (*SD* = 34.22), *t*(27) = -2.699, *p* < .012 VH: *slope* = -5.11 (*SD* = 16.53), *t*(26) = -1.607, *p* < .120	Not reported	No PJ

Nathan et al. (2009): fixed standard condition	18 students	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -34.8, $t(7) = 7.01$, $p < .001$, $R^2 = .891$	*slope* = -200.7, $t(7) = -6.8$, $p < .001$, $R^2 = 0.988$	No PJ
Nuerk et al. (2005)	24 adults	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -3.65, $t(23) = -.89$, $p = .19$ (one-sided), $R^2 = 0.93$	*slope* = -10.13, $t(23) = 1.84$, $p < .05$ (one-sided), $R^2 = 0.96$	No correlation reported
			Repeated-measures regression with both predictors: Categorical: $t(23) = -2.02$, $p < .05$, Continuous: $t(23) = -0.10$, $p = .46$		
Pinto, Pellegrino, Lasaponara, et al. (2021)	15 right brain-damaged adults with left neglect (RBD N+), 17 right brain-damaged adults without left neglect (RBD N-), 15 healthy controls (HC)	1, 2, 3, 4, 6, 7, 8, 9	RBG N+: *slope* = -42.35 (*SD* = 48.51), $t(14) = -3.27$, $p < .01$ RBG N-: *slope* = -27.34 (*SD* = 32.15), $t(16) = -3.51$, $p < .01$ HC: *slope* = -12.53 (*SD* = 22.09), $t(14) = -2.21$, $p < .05$	Not reported	No PJ
Schiller et al. (2016): Experiment 2	24 adults	1, 2, 3, 4, 6, 7, 8, 9 (distance between response keys was manipulated: 6 cm vs. 57 cm vs. 108 cm)	6 cm: *slope* = -11.62, $t(23) = 5.46$, $p < .001$ 57 cm: *slope* = -11.37, $t(23) = 4.37$, $p < .001$ 108 cm: *slope* = -8.34, $t(23) = 3.64$, $p < .002$	Not reported	MC and PJ in between-subjects design

Shaki and Gevers (2011): Experiment 3	46 students	1, 2, 3, 4, 6, 7, 8, 9	*slope* = -5.43, *t*(43) = -2.58, *p* < .01 (one-sided), R^2 = 0.41	Not reported	No PJ
van Dijck and Doricchi (2019)	13 right brain-damaged adults with left neglect (RBD N+), 8 right brain-damaged adults without left neglect (RBD N-), 12 healthy controls (HC)		RBD N+: *slope* = -63.68 (*SD* = 60.81), *t*(11) = -3.63, *p* = 0.004 (one-sided) RBD N-: *slope* = -81.01 (*SD* = 52.24), *t*(7) = -4.39, *p* = 0.003 (one-sided) HC: *slope* = -10.95 (*SD* = 21.85), *t*(11) = -1.74, *p* = 0.055 (one-sided)	Not reported	Spearman *r* between -0.09 and 0.13 with *p* > .480 in each group
van Dijck et al. (2009): Experiment 2 without working memory load	80 students	1, 2, 3, 4, 6, 7, 8, 9	verbal baseline: *slope* = -5.88, *t*(34) = -3.34, *p* < .01 spatial baseline: *slope* = -6.50, *t*(35) = -4.00, *p* < .01	Not reported	Not reported
van Dijck et al. (2012)	10 right brain-damaged adults with left neglect (RBD N+), 7 right brain-damaged adults without	1, 2, 3, 4, 6, 7, 8, 9	RBD N+: *slope* = -70.87 (*SD* = 73.22), RBD N-: *slope* = -56.99 (*SD* = 39.02), HC: *slope* = -12.97 (*SD* = 23.61); all *t*(8, 6, 12) < -1.90, all *p* < .05 (one-sided)	Not reported	Not reported

	left neglect (RBD N-), 12 healthy controls (HC)				
van Galen and Reitsma (2008)	33 7-year-olds, 29 8-year-olds, 27 9-year-olds, 18 adults	1, 2, 3, 4, 6, 7, 8, 9	7-year-olds: *slope* = -47.5, $p < .001$, $R^2 = 0.13$ 8-year-olds: *slope* = -19.6, $p = .007$, $R^2 = 0.04$ 9-year-olds: *slope* = -33.6, $p < .001$, $R^2 = 0.12$ Adults: *slope* = -10.6, $p < .001$, $R^2 = 0.14$	Not reported	No PJ
Weis et al. (2008): Experiment 1	22 students	21, 22, 23, 24, 26, 27, 28, 29, 81, 82, 83, 84, 86, 87, 88, 89 (numbers had to be classified as smaller or larger than the reference number 50)	Unit: *slope* = -0.24, $t(18) = -0.06$, $p = .951$	Decade: *slope* = -14.35, $t(18) = -3.37$, $p = .003$ Unit: *slope* = -66.61, $t(18) = -3.85$, $p = .001$	Not reported
Zorzi et al. (2012)	12 right brain-damaged adults with left neglect (N+), 8 right brain-damaged	1, 2, 3, 4, 6, 7, 8, 9	Not reported	N+: $t(5) = -3.33$, $p < 0.05$ (one-sided) N-: $t(6) = -1.95$, $p < 0.05$ (one-sided)	Not reported

	adults without left neglect (N-)				

Note. This table provides an overview of continuous and categorical number magnitude predictors within repeated-measures regressions for MC used in previous studies. If not specified differently, in all these studies, participants had to classify centrally presented Arabic digits as smaller or larger than the fixed reference of 5 by pressing a left or right response key. Magnitude comparison tasks where two numbers are presented next to each other are not included in this overview. For studies that additionally included parity judgment (PJ), the correlation between the continuous MC-SNARC and the PJ-SNARC slopes is reported (if available). Studies were not included if stimuli were presented in other modalities than visually (e.g., auditorily in Castronovo & Seron, 2007; Beecham et al., 2009; Weis et al., 2015), if responses were not given with the left and right index fingers using one left and one right response key (e.g., unimanual responses in Cipora, 2014; Riello & Rusconi, 2011; vocal responses in Leth-Steensen & Citta, 2016; four response keys in Santens & Gevers, 2008; fields with “left” and “right” labels on touch screen in Gevers et al., 2010), if the response rule was changing within blocks (e.g., response-to-key assignment announced before each trial in Basso Moro et al., 2018; response rule depending on stimulus in Notebaert et al., 2006; Pinto et al., 2021), if too few stimuli to differentiate the continuous from a categorical SNARC shape (e.g., 1, 2, 8, and 9 in Fischer et al., 2016; Imbo et al., 2012; Weis et al., 2018; 1, 4, 6, and 9 in Santens & Gevers, 2008) were used, if the stimulus set also included negative numbers (Fischer & Rottmann, 2005; Shaki & Petrusic, 2005), or if a distractor was included (e.g., Chinese character meaning “small” or “large” in the background in Nan et al., 2022). If not specified otherwise, all reported tests were two-sided.

References for Appendix A

Bachot, J., Gevers, W., Fias, W., & Roeyers, H. (2005). Number sense in children with visuospatial disabilities: Orientation of the mental number line. *Psychology Science, 47*(1), 172–183.

Bae, G. Y., Choi, J. M., Cho, Y. S., & Proctor, R. W. (2009). Transfer of magnitude and spatial mappings to the SNARC effect for parity judgments. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*(6), 1506–1521. https://doi.org/10.1037/a0017257

Basso Moro, S., Dell'Acqua, R., & Cutini, S. (2018). The SNARC effect is not a unitary phenomenon. *Psychonomic Bulletin & Review, 25*(2), 688–695. https://doi.org/10.3758/s13423-017-1408-3

Beecham, R., Reeve, R. A., & Wilson, S. J. (2009). Spatial representations are specific to different domains of knowledge. *PLOS ONE, 4*(5), e5543. https://doi.org/10.1371/journal.pone.0005543

Bull, R., Marschark, M., & Blatto-Vallee, G. (2005). SNARC hunting: Examining number representation in deaf students. *Learning and Individual Differences, 15*(3), 223–236. https://doi.org/10.1016/j.lindif.2005.01.004

Bulut, M., Roth, L., Bahreini, N., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2025). One direction? Cultural aspects of the mental number line beyond reading direction. Psychological Research, 89, Article 37. https://doi.org/10.1007/s00426-024-02038-4

Castronovo, J., & Seron, X. (2007). Semantic numerical representation in blind subjects: The role of vision in the spatial format of the mental number line. *The Quarterly Journal of Experimental Psychology, 60*(1), 101–119. https://doi.org/10.1080/17470210600598635

Cheung, C.-N., Ayzenberg, V., Diamond, R. F. L., Yousif, S., & Lourenco, S. F. (2015). *Probing the mental number line: A between-task analysis of spatial-numerical associations.* Proceedings of the 37th Annual Meeting of the Cognitive Science Society., 357–362.

Cipora, K. (2014). *Between-task consistency of the SNARC effect.* XXXIInd European Workshop on Cognitive Neuropsychology, Bressanone, Italy. Retrieved from https://osf.io/nx6vq/

Cipora, K., Hohol, M., Nuerk, H.-C., Willmes, K., Brożek, B., Kucharzyk, B., & Nęcka, E. (2016). Professional mathematicians differ from controls in their spatial-numerical associations. *Psychological Research*, *80*(4), 710–726. https://doi.org/10.1007/s00426-015-0677-6

Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Goebel, S. M., Willmes, K., Pesenti, M., Schiltz, C., & Nuerk, H.-C. (2019). *A minority pulls the sample mean: On the individual prevalence of robust group-level cognitive phenomena – the instance of the SNARC effect* [preprint]. Open Science Framework. https://doi.org/10.31234/osf.io/bwyr3

Deng, Z., Chen, Y., Zhu, X., & Li, Y. (2017). The effect of working memory load on the SNARC effect: Maybe tasks have a word to say. *Memory & Cognition, 45*(3), 428–441. https://doi.org/10.3758/s13421-016-0676-x

Didino, D., Breil, C., & Knops, A. (2019). The influence of semantic processing and response latency on the SNARC effect. *Acta Psychologica, 196*, 75–86. https://doi.org/10.1016/j.actpsy.2019.04.008

Fattorini, E., Pinto, M., Rotondaro, F., & Doricchi, F. (2015). Perceiving numbers does not cause automatic shifts of spatial attention. *Cortex, 73*, 298–316. https://doi.org/10.1016/j.cortex.2015.09.007

Fischer, M. H., & Rottmann, J. (2005). Do negative numbers have a place on the mental number line. *Psychology Science, 47*(1).

Fischer, U., Moeller, K., Class, F., Huber, S., Cress, U., & Nuerk, H.-C. (2016). Dancing with the SNARC: Measuring spatial-numerical associations on a digital dance mat. *Canadian Journal of Experimental Psychology, 70*(4), 306–315. https://doi.org/10.1037/cep0000084

Fitousi, D., Shaki, S., & Algom, D. (2009). The role of parity, physical size, and magnitude in numerical cognition: The SNARC effect revisited. *Attention, Perception & Psychophysics, 71*(1), 143–155. https://doi.org/10.3758/APP.71.1.143

Georges, C., Hoffmann, D., & Schiltz, C. (2017). How and why do number-space associations co-vary in implicit and explicit magnitude processing tasks? *Journal of Numerical Cognition, 3*(2), 182–211. https://doi.org/10.5964/jnc.v3i2.46

Gevers, W., Caessens, B., & Fias, W. (2005). Towards a common processing architecture underlying Simon and SNARC effects. *European Journal of Cognitive Psychology, 17*(5), 659–673. https://doi.org/10.1080/09541440540000112

Gevers, W., Santens, S., Dhooge, E., Chen, Q., van den Bossche, L., Fias, W., & Verguts, T. (2010). Verbal-spatial and visuospatial coding of number-space interactions. *Journal of Experimental Psychology: General, 139*(1), 180–190. https://doi.org/10.1037/a0017688

Gevers, W., Verguts, T., Reynvoet, B., Caessens, B., & Fias, W. (2006). Numbers and space: A computational model of the SNARC effect. *Journal of Experimental Psychology: Human Perception and Performance, 32*(1), 32–44. https://doi.org/10.1037/0096-1523.32.1.32

Han, M., Mao, X., Cai, M., Jia, X., & Guo, C. (2017). The effect of positive and negative signs on the SNARC effect in the magnitude judgment task. *Acta Psychologica Sinica, 49*(8), 995–1008. https://doi.org/10.3724/SP.J.1041.2017.00995

Herrera, A., Macizo, P., & Semenza, C. (2008). The role of working memory in the association between number magnitude and space. Acta Psychologica, 128(2), 225–237. https://doi.org/10.1016/j.actpsy.2008.01.002

Hoffmann, D., Hornung, C., Martin, R., & Schiltz, C. (2013). Developing number-space associations: SNARC effects using a color discrimination task in 5-year-olds. *Journal of Experimental Child Psychology, 116*(4), 775–791. https://doi.org/10.1016/j.jecp.2013.07.013

Hohol, M., Willmes, K., Nęcka, E., Brożek, B., Nuerk, H.-C., & Cipora, K. (2020). Professional mathematicians do not differ from others in the symbolic numerical distance and size effects. *Scientific Reports, 10*(1), 11531. https://doi.org/10.1038/s41598-020-68202-z

Hubbard, E. M., Ranzini, M., Piazza, M., & Dehaene, S. (2009). What information is critical to elicit interference in number-form synaesthesia? *Cortex, 45*(10), 1200–1216. https://doi.org/10.1016/j.cortex.2009.06.011

Imbo, I., Brauwer, J. de, Fias, W., & Gevers, W. (2012). The development of the SNARC effect: Evidence for early verbal coding. *Journal of Experimental Child Psychology, 111*(4), 671–680. https://doi.org/10.1016/j.jecp.2011.09.002

Ito, Y., & Hatta, T. (2004). Spatial structure of quantitative representation of numbers: Evidence from the SNARC effect. *Memory & Cognition, 32*(4), 662–673. https://doi.org/10.3758/BF03195857

Leth-Steensen, C., & Citta, R. (2016). Bad-good constraints on a polarity correspondence account for the spatial-numerical association of response codes (SNARC) and markedness association of response codes (MARC)

effects. *Quarterly Journal of Experimental Psychology, 69*(3), 482–494. https://doi.org/10.1080/17470218.2015.1055283

Lohmann, J., Schroeder, P. A., Nuerk, H.-C., Plewnia, C., & Butz, M. V. (2018). How deep is your SNARC? Interactions between numerical magnitude, response hands, and reachability in peripersonal space. *Frontiers in Psychology, 9*, Article 622. https://doi.org/10.3389/fpsyg.2018.00622

Mourad, A., & Leth-Steensen, C. (2017). Spatial reference frames and SNARC. *Journal of Cognitive Psychology, 29*(2), 113–128. https://doi.org/10.1080/20445911.2016.1249483

Nan, W., Yan, L., Yang, G., Liu, X., & Fu, S. (2022). Two processing stages of the SNARC effect. *Psychological Research, 86*(2), 375–385. https://doi.org/10.1007/s00426-021-01506-5

Nathan, M. B., Shaki, S., Salti, M., & Algom, D. (2009). Numbers and space: Associations and dissociations. *Psychonomic Bulletin & Review, 16*(3), 578–582. https://doi.org/10.3758/PBR.16.3.578

Notebaert, W., Gevers, W., Verguts, T., & Fias, W. (2006). Shared spatial representations for numbers and space: The reversal of the SNARC and the Simon effects. *Journal of Experimental Psychology Human Perception & Performance, 32*(5), 1197–1207. https://doi.org/10.1037/0096-1523.32.5.1197

Nuerk, H.-C., Bauer, F., Krummenacher, J., Heller, D., & Willmes, K. (2005). The power of the mental number line: How the magnitude of unattended numbers affects performance in an Eriksen task. *Psychology Science, 47*(1), 34–50. https://psycnet.apa.org/record/2005-11470-005

Pinto, M., Pellegrino, M., Lasaponara, S., Scozia, G., D'Onofrio, M., Raffa, G., Nigro, S., Arnaud, C. R., Tomaiuolo, F., & Doricchi, F. (2021). Number space is made

by response space: Evidence from left spatial neglect. *Neuropsychologia, 154*, 107773. https://doi.org/10.1016/j.neuropsychologia.2021.107773

Pinto, M., Pellegrino, M., Marson, F., Lasaponara, S., Cestari, V., D'Onofrio, M., & Doricchi, F. (2021). How to trigger and keep stable directional Space-Number Associations (SNAs). *Cortex, 134*, 253–264. https://doi.org/10.1016/j.cortex.2020.10.020

Riello, M., & Rusconi, E. (2011). Unimanual SNARC effect: Hand matters. *Frontiers in Psychology, 2*, Article 372. https://doi.org/10.3389/fpsyg.2011.00372

Santens, S., & Gevers, W. (2008). The SNARC effect does not imply a mental number line. *Cognition, 108*(1), 263–270. https://doi.org/10.1016/j.cognition.2008.01.002

Schiller, F., Eloka, O., & Franz, V. H. (2016). Using key distance to clarify a theory on the SNARC. *Perception, 45*(1-2), 196–221. https://doi.org/10.1177/0301006615616754

Shaki, S., & Gevers, W. (2011). Cultural characteristics dissociate magnitude and ordinal information processing. *Journal of Cross-Cultural Psychology, 42*(4), 639–650. https://doi.org/10.1177/0022022111406100

Shaki, S., & Petrusic, W. M. (2005). On the mental representation of negative numbers: Context-dependent SNARC effects with comparative judgments. *Psychonomic Bulletin & Review, 12*(5), 931–937. https://doi.org/10.3758/BF03196788

Van Dijck, J.-P., & Doricchi, F. (2019). Multiple left-to-right spatial representations of number magnitudes? Evidence from left spatial neglect. *Experimental Brain Research, 237*(4), 1031–1043. https://doi.org/10.1007/s00221-019-05483-5

Van Dijck, J.-P., Gevers, W., & Fias, W. (2009). Numbers are associated with different types of spatial information depending on the task. *Cognition, 113*(2), 248–253. https://doi.org/10.1016/j.cognition.2009.08.005

Van Dijck, J.-P., Gevers, W., Lafosse, C., & Fias, W. (2012). The heterogeneous nature of number-space interactions. *Frontiers in Human Neuroscience, 5*, Article 182. https://doi.org/10.3389/fnhum.2011.00182

Van Galen, M. S., & Reitsma, P. (2008). Developing access to number magnitude: A study of the SNARC effect in 7- to 9-year-olds. *Journal of Experimental Child Psychology, 101*(2), 99–113. https://doi.org/10.1016/j.jecp.2008.05.001

Weis, T., Estner, B., Krick, C. M., Reith, W., & Lachmann, T. (2015). SNARC meets SPARC in fMRI--Interdependence of compatibility effects depends on semantic content. *Neuropsychologia, 77*, 331–338. https://doi.org/10.1016/j.neuropsychologia.2015.09.018

Weis, T., Nuerk, H.-C., & Lachmann, T. (2018). Attention allows the SNARC effect to operate on multiple number lines. *Scientific Reports, 8*(1), 13778. https://doi.org/10.1038/s41598-018-32174-y

Weis, T., Theobald, S., Schmitt, A., van Leeuwen, C., & Lachmann, T. (2018). There's a SNARC in the Size Congruity Task. *Frontiers in Psychology, 9*, Article 1978. https://doi.org/10.3389/fpsyg.2018.01978

Zorzi, M., Bonato, M., Treccani, B., Scalambrin, G., Marenzi, R., & Priftis, K. (2012). Neglect impairs explicit processing of the mental number line. *Frontiers in Human Neuroscience, 6*, 125. https://doi.org/10.3389/fnhum.2012.00125

Appendix B

Non-preregistered exploratory analyses

In addition to the exploratory analyses described in the stage-1 Registered Report, we took advantage of the large amount of collected data and ran several more exploratory analyses.

Exploratory 8

We tested whether responses were faster in the second block than in the first block due to learning effects in two-sided paired *t*-tests separately within each task. Evidence strongly supported this for MC (BF_{10} = 1341.81, Cohen's d = 0.19, with $M_{RT\ first\ block}$ = 531 ms, SD = 104 ms, and $M_{RT\ second\ block}$ = 523 ms, SD = 100 ms), whereas evidence was moderately against this expectation for PJ (BF_{10} = 0.23, Cohen's d = 0.07, with $M_{RT\ first\ block}$ = 579 ms, SD = 103 ms, and $M_{RT\ second\ block}$ = 576 ms, SD = 104 ms).

Exploratory 9

Similarly, we investigated whether responses were faster in the second as compared to the first half of each block in two-sided paired *t*-tests separately within each task. This time, as opposed to Exploratory 8, evidence was strongly against such an observation for MC (BF_{10} = 0.05, Cohen's d = -0.02, with $M_{RT\ first\ half}$ = 526 ms, SD = 100 ms, and $M_{RT\ second\ half}$ = 527 ms, SD = 100 ms), but in favor of it for PJ (BF_{10} = 139.97, Cohen's d = 0.16, with $M_{RT\ first\ half}$ = 579 ms, SD = 102 ms, and $M_{RT\ second\ half}$ = 574 ms, SD = 101 ms).

Exploratory 10

To investigate whether an Odd effect could be replicated in each task, we tested the RT differences (odd minus even) against 0 in a two-sided paired *t*-test. Surprisingly, we found strong evidence for a reversed Odd effect in MC (BF_{10} = $5.48 * 10^{72}$, Cohen's

d = -0.87, with $M_{RT\ odd}$ = 522 ms, SD = 99 ms, and $M_{RT\ even}$ = 532 ms, SD = 101 ms). In contrast, we found strong evidence for a typical Odd effect in PJ (BF_{10} = 6.47 * 10^{29}, Cohen's d = 0.52, with $M_{RT\ odd}$ = 582 ms, SD = 102 ms, and $M_{RT\ even}$ = 572 ms, SD = 100 ms).

Exploratory 11

Next, we checked SNARC-compatibility effects by testing whether responses were faster in SNARC-compatible than in SNARC-incompatible trials using two-sided paired *t*-tests separately in each task. Note that the trials' SNARC-compatibility was grouped blockwise in MC, but intermixed in PJ. We found strong evidence for the expected SNARC-compatibility effect both in MC (BF_{10} = 11225372, Cohen's d = 0.26, with $M_{RT\ SNARC\text{-}incompatible}$ = 532 ms and $M_{RT\ SNARC\text{-}compatible}$ = 522 ms) and in PJ (BF_{10} = 6.13 * 10^{26}, Cohen's d = 0.49, with $M_{RT\ SNARC\text{-}incompatible}$ = 581 ms and $M_{RT\ SNARC\text{-}compatible}$ = 573 ms).

Exploratory 12

In a similar vein, we investigated MARC-compatibility effects by testing whether responses were faster in MARC-compatible than in MARC-incompatible trials using two-sided paired *t*-tests separately in each task. Note that the trials' MARC-compatibility was intermixed in MC, but grouped blockwise in PJ. Strong evidence against such an effect was found in MC (BF_{10} = 0.05, Cohen's d = 0.02, with $M_{RT\ MARC\text{-}incompatible}$ = 527 ms and $M_{RT\ MARC\text{-}compatible}$ = 526 ms) as well as in PJ (BF_{10} = 0.05, Cohen's d = 0.02, with $M_{RT\ MARC\text{-}incompatible}$ = 578 ms, and $M_{RT\ MARC\text{-}compatible}$ = 577 ms).

Exploratory 13

We also checked whether the Numerical Size Effect was correlated with overall RT. Evidence was moderately against such a correlation both in MC (r = 0.05, BF_{10} = 0.23) and in PJ (r = -0.03, BF_{10} = 0.12).

Exploratory 14

We re-calculated the PJ-SNARC effect while only considering post-SNARC-compatible trials and excluding all post-SNARC-incompatible trials (as suggested by Moretti et al., 2025). As in Hypothesis 2, a two-sided Bayesian paired *t*-test revealed strong evidence for a better fit of the continuous model PJ-1 (R^2 = 0.342, R^2_{logit} = -0.653) than for the categorical model PJ-2 (R^2 = 0.312, R^2_{logit} = -0.79) with BF_{10} = 74.41. Notably, the SNARC effect size was much larger than when considering both post-SNARC-compatible and post-SNARC-incompatible trials (slope = -11.01 ms, *SD* = 8.65 ms, *d* = -1.27).

Exploratory 15

In addition to Exploratory 14, we also checked whether the split-half reliability improved when only considering post-SNARC-compatible trials and excluding all post-SNARC-incompatible trials (as suggested by Moretti et al., 2025). Evidence for the split-half reliability of the continuous PJ-SNARC was strong with an estimated correlation of *r* = 0.28 (Spearman-Brown corrected estimate *r* = 0.44; BF_{10} = 2.97 * 10^9), which is similar to when both post-SNARC-compatible and post-SNARC-incompatible trials were considered.

Exploratory 16

To explore whether the MARC effect in PJ differed between left- and right-handers, we ran a two-sided independent-samples *t*-tests against 0. Moderate evidence against a difference (BF_{10} = 0.15) between the data from 530 right-handers (*d* = -0.02, slope = -1.67 ms, *SD* = 91.88 ms) and the data from 56 left-handers (*d* = 0.00, slope = 0.03 ms, *SD* = 80.18 ms) was found.

Exploratory 17

To investigate whether the shape of the MC-SNARC was mixed, having both continuous and categorical aspects, we ran Exploratory 17 in addition to Exploratory

3. We computed the continuous regression slopes within small and large numbers separately and tested them against 0 in two-sided one-sample t-tests. Data revealed strong evidence for negative slopes different from 0 in both small numbers (slope = -5.13 ms, *SD* = 13.52 ms, *d* = -0.38, BF_{10} = 2.21 * 10^{16}) and large numbers (slope = -2.41 ms, *SD* = 14.54 ms, *d* = -0.17, BF_{10} = 165.99).

References for Appendix B

Moretti, L., Koch, I., Hornjak, R., & von Bastian, C. C. (2025). Quality over quantity: Focusing on high-conflict trials to improve the reliability and validity of attentional control measures. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* https://doi.org/10.1037/xlm0001466

A3. Study 3

A3.1. Publication of Study 3

Citation:
Bulut, M., Roth, L., Bahreini, N., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2025). One direction? Cultural aspects of the mental number line beyond reading direction. *Psychological Research, 89*, Article 37. https://doi.org/10.1007/s00426-024-02038-4

Copyright:
© 2025 Bulut, Roth, Bahreini, Cipora, Reips, Nuerk (Open Access)
This article is published under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, provided the original authors and source are credited.

Psychological Research (2025) 89:37
https://doi.org/10.1007/s00426-024-02038-4

RESEARCH

One direction? Cultural aspects of the mental number line beyond reading direction

Merve Bulut[1] · Lilly Roth[2] · Narjes Bahreini[2] · Krzysztof Cipora[3] · Ulf Dietrich Reips[4] · Hans-Christoph Nuerk[2]

Received: 28 June 2024 / Accepted: 15 October 2024
© The Author(s) 2024

Abstract

Spatial-Numerical Associations (SNAs) refer to the demonstrations of spatial processing of numbers. The *Mental Number Line* (MNL) is a representation model describing numbers as aligning left-to-right (LR) and was suggested to account for directional biases in participants' responses during numerical tasks. One common behavioral demonstration of this is the *Spatial-Numerical Associations of Response Codes* (SNARC) effect, which describes faster left-/right-hand responses to smaller/larger numbers, respectively. The MNL, and, consequently, directional SNAs, show variabilities across different cultures. Reading direction is considered to be the main factor in explaining these differences. In line with this, individuals with right-to-left (RL) reading habits show a weaker or even reverse SNARC effect. In the present study, we investigated whether SNAs are influenced not only by reading direction, but also by cultural directional preferences such as drawing lines, arranging objects, imagining objects (i.e., rightward or leftward facing), or representing events in time (i.e., mentally representing the past/future on the left/right, respectively). To test this hypothesis, we measured the cultural directional preferences and the SNARC effect across three cultures in an online setup; German, Turkish, and Iranian. LR preferences in the Cultural Directional Preferences Questionnaire were most prominent in German participants, intermediate in Turkish participants, and least prominent in Iranian participants. In line with this, the LR SNARC effect was strongest in German, intermediate in Turkish, and weakest (but not RL) in Iranian culture. These findings suggest that cultural directional preferences are involved in the emergence of adult SNAs in addition to the reading direction.

Introduction

Numeracy is a crucial skill for understanding the external world and has many applications in everyday life. Therefore, it is essential to understand how humans process numbers and create a meaningful mental representation of them. One core concept related to elementary number processing is the strong cognitive link between numerical magnitude and space. *Spatial-Numerical Associations* (SNAs) are a group of behavioral and neural demonstrations indicating that numerical information is spatially represented (for a review see Cipora et al., 2018).

One of the most reported behavioral demonstrations of SNAs is the *Spatial-Numerical Association of Response Codes* (SNARC; Dehaene et al., 1993). In a given numerical range, smaller numbers are responded to faster with left-side buttons and larger numbers with right-side buttons. The SNARC effect is thought to be a result of the left-to-right (LR) orientation of the *Mental Number Line* (MNL), on which smaller numbers are mapped on the left and larger ones on the right (Dehaene et al., 1993; for alternative accounts see Gevers et al., 2010; van Dijck & Fias, 2011; Schroeder et al., 2017; Cipora et al., 2020).

One line of research focuses on whether this close link between space and number has a biological basis or is culturally acquired. Animal studies suggest that LR spatial processing of magnitudes is innate by demonstrating LR processing of magnitudes in newly born chicks (Rugani et al., 2015) and macaque monkeys (Adachi, 2014; Drucker

✉ Hans-Christoph Nuerk
hc.nuerk@uni-tuebingen.de

[1] Department of Psychology, Yaşar University, Izmir, Turkey

[2] Department of Psychology, University of Tuebingen, Schleichstrasse 4, 72072 Tuebingen, Germany

[3] Department of Mathematics Education, Loughborough University, Loughborough, UK

[4] Department of Psychology, University of Konstanz, Constance, Germany

Published online: 23 December 2024

& Brannon, 2014). This view is also supported by evidence from studies with preliterate infants. Specifically, infants seem to process increasing magnitudes in LR fashion (e.g., de Hevia et al., 2014; Bulf et al., 2016, Di Giorgio et al., 2019; for a review see McCrink & Opfer, 2014). These studies show that the spatial processing of magnitudes appears very early in development. However, since these studies only include Western infants born and raised in the LR-oriented cultural environment, their behaviors may be highly influenced by their noticing of environmental clues (see McCrink & de Hevia, 2018; Patro et al., 2016a for discussions). Importantly, —whether or not these tendencies are innately present—it is well-demonstrated that culture can shape SNAs, as we will see in the following section. The present paper focuses on how cultural directionalities beyond reading/writing direction (referred to as reading direction in the text hereafter) can influence the spatial processing of numbers.

The influence of reading direction on SNAs

Reading direction was proposed as the source of SNAs soon after the SNARC effect was first observed. Already in the first systematic investigation of the SNARC effect, Dehaene et al. (1993; Exp.7) reported that it was absent among Iranian participants who live in France. The right-to-left (RL) reading direction of the Farsi (also known as Persian) language was suggested as an explanation. The results were similar when Arabic and Eastern Arabic (typically used in Farsi) numbers were used as stimuli. Dehaene et al. concluded that the absence of the SNARC effect in Farsi speakers could be related to the fact they were using both LR (French) and RL (Farsi) reading directions. An additional analysis revealed a significant relationship between the time participants spent in France and the SNARC effect; the longer they spent in France, the stronger the SNARC effect.

Shaki et al., (2009) examined the effect of reading direction on the SNARC effect in a cross-cultural setting more systematically. On top of reading directions for text, they considered whether numbers in a given language are written LR or RL. These directions do not match in some languages, including Farsi and Hebrew (text is written from right to left, but numbers are written from left to right). They recruited three samples: Canadians (LR of both words and Arabic numbers), Israelis (RL reading of words and LR reading of Arabic numbers), and Palestinians (RL reading of both words and Eastern Arabic numbers). Canadians revealed a regular SNARC effect, Palestinians a reverse SNARC effect (i.e., faster left-/right responses to large/small numbers, respectively), and Israelis no effect. Shaki et al. (2009) also attributed the differences in the SNARC effect to reading habits (see also Zebian, 2005). The null effect in Hebrew speakers was attributed to the inconsistent reading direction of text (RL) and numbers (LR). However, one must keep in mind that in these early studies, the statistical power was poor for detecting small effect sizes due to small samples (between 11 and 20 participants, the minimum detectable effect size would have been $d = 0.94$ to $d = 0.66$ for 0.80 power, respectively) and poor measurement precision resulting from insufficient numbers of repetitions per combination of number and response-to-key assignment (between 8 and 16 repetitions), whereas at least 20 repetitions are recommended (see Cipora & Wood, 2017). At the same time, several subsequent studies demonstrated a regular SNARC effect at the group level in RL readers (Feldman et al., 2019; Hochman et al., 2024; Shaki & Gevers, 2011; Zohar-Shai et al., 2017; Supplementary Material of Cipora et al., 2019a).

The connection between reading direction on the SNARC effect is further supported by studies testing bilingual individuals with both LR (Russian) and RL (Hebrew) reading habits. The SNARC effect can alter even within a very short time after reading a Russian (regular SNARC) or Hebrew (no SNARC) paragraph (Shaki & Fischer, 2024) or even a single Russian or Hebrew number word (Fischer et al., 2009). This suggests that reading direction (and associated cultural influences) not only has long-term memory-related but also short-term priming-like influences on the SNARC effect. At the same time, the conclusion that reading manipulation can swiftly alter the SNARC has not always been supported. For instance, Pitt and Casasanto (2020) examined the causal influence of short-time reading activity on the SNARC effect by giving English-speaking participants a text in their native language either with a standard or mirror-reversed orthography. The SNARC effects after reading in a standard and reverse fashion were similar, showing no evidence of the influence of reading direction on the SNARC effect.

To sum up, while the impact of the reading direction of the SNARC is still far from fully resolved, studies looking into this effect have provided some hints about the role of other directional habits in shaping the SNARC effect. At the same time, the direction in which numbers are written is more nuanced and does not fully allow testing this hypothesis. It is important to differentiate writing numbers in a text line and the numeration systems. Although the Eastern Arabic numbers used by Palestinian Arabic speakers and Farsi speakers are written RL in line organization (i.e., one writes digits from one to five starting from right), they are written LR in the numeration system (i.e., multi-digit numbers have the highest power to ten on the left and the unit on the right as in Western languages including English; see Lopiccolo & Chang, 2021; Rashidi-Ranjbar et al., 2014). Hence, differences between Hebrew and Arabic do not include the direction of their numeration systems (i.e., they are both LR).

The influence of cultural directional preferences on SNAs

Although the evidence reviewed above suggests that reading direction influences SNAs, it does not indicate that reading direction is the only cultural factor involved in their emergence. Patro et al. (2016b) proposed an explanation of how engaging in actions that include a spatial trajectory (such as reading a text, or horizontal ordering of items) can influence numerical magnitude processing. Accordingly, the direction of action corresponds with the time passing, spatial distance from the starting point, and the amount of objects involved. For instance, when observing the activity of arranging books on a shelf in an LR fashion, one can realize that as the action moves towards the right, more time passes, the space covered in the shelf increases, and more books accumulate. Therefore, engaging in spatially oriented behaviors or observing them may enhance the spatial processing of magnitudes.

In line with this, evidence from preschool children studies shows that SNAs are established before reading skills emerge. Hoffmann et al. (2013) found a significant regular SNARC effect with 5.5-year-old children in a color classification task of Arabic numbers (but only when a magnitude classification task of Arabic numbers had been previously performed). Moreover, with a non-symbolic magnitude comparison task, Ebersbach et al. (2014) showed that SNAs are prevalent before formal education (see also Patro & Haman, 2012). Similarly, Shaki et al. (2012) reported culture-specific object-counting preferences among preschool children consistent with the prominent reading direction (LR in British, RL in Palestinian, and no dominant preferences in Israeli preschoolers; see also Patro et al., 2015).

To explain how SNAs might be constructed by the cultural environment in preschool children, Patro et al. (2016b) proposed the *implicit instruction account*. The account suggests that SNAs in preliterate children are not necessarily innate, but can emerge from implicit learning of culture-specific directional preferences acquired by observing and interacting with parents and their social environment. To support their hypothesis, Patro et al. (2016b) trained 3- to 4-year-old preliterate children with a non-numerical task that involves spatial finger movements (LR or RL) on a touch screen to move a frog across a pond. After the training, the LR group showed a regular SNARC effect and the RL group showed a reversed SNARC effect in a numerical comparison task. These findings fit well with Patro et al. (2016a, 2016b)'s explanation of the emergence of SNAs mentioned above. Visuomotor training in a specific direction (LR or RL) enhances the spatial trajectory of longer duration, larger numerosity, and larger distance, compatible with the training direction (see also McCrink & de Hevia, 2018). Therefore, these findings suggest that directional visio-motor behaviors can establish (at least temporarily) SNAs. Göbel et al. (2018) demonstrated a similar effect of the observed reading on the counting preferences. 3- to 5-year-old English- and Arabic-speaking preliterate children changed their counting preference in line with the direction of the observed storybook reading (LR or RL). Shaki and Fischer (2023) provided further evidence for the link between counting preference and attentional SNARC in adults.

In a similar vein, Nuerk et al. (2015) proposed several mechanisms to explain how certain behaviors can influence the emergence of SNAs among preliterate children. The emergence of SNAs may result not only from observing adult reading behavior and pretending to read/write, but also from "dominant attentional-directional preferences in a society, not directly related to reading direction" (Nuerk et al., 2015, p. 3) such as drawing lines, arranging items, imagining objects (i.e., rightward or leftward facing), or representing events in time (i.e., mentally representing the past and the future on the left or right; see also Patro et al., 2016a). Here, we refer to these behaviors as cultural directional preferences, and for clarity, we separate cultural directionalities as (i) reading direction (which is obligatory in society) and (ii) cultural directional preferences unrelated to reading direction. Previous studies investigating culture's influence on adult SNAs have focused on reading direction, overlooking cultural directional preferences, even though, as summarized above, studies with preliterate children suggest that directional actions other than reading are also involved.

The postulate that factors other than reading direction are also responsible for the SNARC effect is considered in the recent *CORrelations in Experience (CORE) principle* (Pitt & Casasanto, 2020), stating that different experiences such as reading and finger counting should selectively influence different abstract representations such as MNL. Based on this, reading activity highly correlates with time and space (more time passes by when the reader moves in the text in LR or RL) but not with number and space unless the text includes numbers; therefore, there is no need to expect that there will be a causal relationship between text and MNL directions.

To summarize, cultural directional preferences can influence SNAs. This evidence comes mainly from studies with preliterate children, where cultural directional preferences seem to impact SNAs even before reading skills are acquired. In the following, we will outline the evidence pertaining to whether cultural directional preferences could be also influential in adult SNAs.

Reading direction versus cultural directional preferences

The theoretical accounts proposed to emphasize the influence of cultural directional preferences were formulated to explain the SNAs that emerge in preliterate children in the process of development. One interesting question is whether cultural directional preferences still influence SNAs in adults. One possibility is that the effects observed among preschool children result from a transient phase in which children use spatial cues in their social environment to process magnitudes. The most common and consistent cultural directionality is reading, and therefore, reading direction would be the prominent factor determining SNAs in adulthood. However, there is another possibility, that as well as reading direction, cultural directional preferences also have an influence.

It is difficult to test whether cultural directional preferences are also involved in the emergence of SNAs in adulthood because these are frequently consistent with the reading direction of that culture (Chokron & de Agostini, 1995; Faghihi et al., 2019; Maass & Russo, 2003; Rinaldi et al., 2014; Schild et al., 2022 Vaid, 1995 for a review see Faghihi & Vaid, 2023). At first sight, it seems unlikely that there are cultures and individuals who might have opposite cultural directional preferences and reading directions. However, Turkish culture may provide this remarkable opportunity to dissociate long-term cultural directional preferences from current reading habits. In the following paragraph, we detail how this unique condition might have emerged.

The official alphabet of the Turkish language was a combination of Arabic-Farsi written RL until 1928 for around six hundred years (during the Ottoman Empire). In 1923, the modern Turkish Republic was founded and there were reforms in several areas. One of these revolutions involved replacing the Arabic-Farsi alphabet with the Latin alphabet in 1928 (Yılmaz, 2011). Even though modern Turkish and the old Ottoman Turkish languages are almost identical, after the reform, the Turkish language was written with the Latin alphabet in an LR fashion. We suggest that this switch from RL to LR direction in the script may create inconsistencies between the reading habits and cultural directional preferences in Turkish culture, because directional preferences do not change automatically with script change (see e.g., Nuerk et al., 2015, for children). It is possible that they survived to the present due to longstanding traditions of Middle Eastern culture, even apart from the previous RL reading direction. For instance, the Islamic religion, still very prominent in Turkey, recommends initiating actions from the right side. Other cultural preferences such as counting, drawing preferences of objects, and ordering preferences might also have survived, despite the change in reading directions. We investigated this issue by using a questionnaire designed to measure cultural directional preferences.

If there are mixed cultural directionalities in Turkish culture (i.e., LR reading and RL directional preferences), then we would expect weaker LR SNAs compared to a typical LR culture in which cultural directional preferences covary with the reading direction. Confirming this expectation, Bulut et al. (2023) reported no significant SNARC effect in a Turkish sample.

The present study

In the present study, we aimed to test the *cultural directionality hypothesis:* We suggest that SNAs are influenced by the cultural environment, including reading habits and many other cultural directional preferences in daily life. These two factors are usually consistent, and thus, indistinguishable, but this might not be the case in Turkish culture due to the script direction change. Some cultural directional preferences might still be RL. Consequently, the possible inconsistencies of directionalities might weaken the LR SNAs in Turkish culture.

To test the cultural directionality hypothesis, we designed a cross-cultural study comparing the effect of reading direction and cultural directional preferences on SNAs. In the existing cross-cultural SNA studies, these two elements are confounded. To disentangle cultural directional preferences from reading direction, we recruited samples from three distinct cultures. We compared directional SNAs in a Turkish sample (hypothetically mixed culture, in which the reading direction currently is LR but cultural directional preferences mostly RL) with a German (predominantly LR) and an Iranian (predominantly RL) sample. The aim of the comparison of the German and Turkish samples was to reveal any influences of cultural directional preferences since reading direction is LR in both cultures. Furthermore, comparing the Iranian and Turkish samples would reveal the influence of reading direction by reducing the involvement of cultural directional preferences, since we expect these to be similarly RL in both (see Fig. 1).

To test the cultural directionality hypothesis, we also aimed to check whether cultural directional preferences in the Turkish sample are, as hypothesized, RL or mixed, with a newly designed Cultural Directional Preferences Questionnaire (CDPQ), introduced below. We expected cultural directional preferences in Turkish participants to be more RL compared to a predominantly LR culture (i.e., German) and more LR compared to a predominantly RL culture (i.e., Iranian).

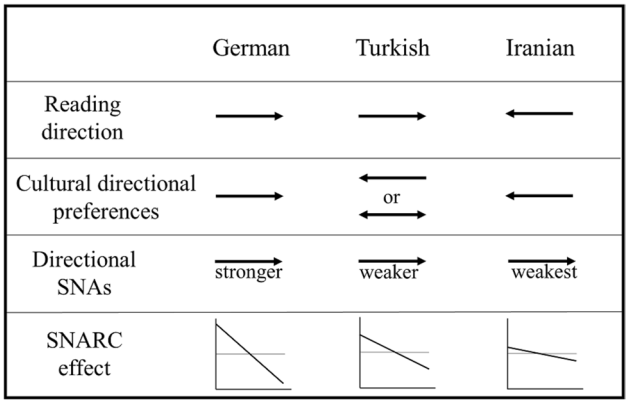


Fig. 1 The depiction of the cultural directionality hypothesis which predicts that both reading direction and cultural directional preferences are involved in the emergence of SNAs. Turkish culture is expected to be mixed regarding reading direction and cultural directional preferences. Therefore, we expected SNAs (specifically the SNARC effect) to be less prominent than a consistent LR culture (i.e., German). Also, we expected the SNARC effect to be more prominent in Turkish than a consistent RL culture (i.e., Iranian) based on the reading direction influence. A steeper negative slope represents a stronger SNARC effect

Importantly, "predominantly LR/RL cultures" does not imply that every individual or every behavior of individuals in these samples is in line with certain directions consistently. On the contrary, there is a wealth of evidence for variance in various forms of directional SNAs within cultures. For instance, only 35–45% of participants in Western samples show a reliable regular SNARC effect and 5–10% even show a reliable reverse SNARC effect (Cipora et al., 2019a, Wood et al., 2008; see also Table 3 for the prevalence of SNARC across cultures in the present study). Moreover, the SNARC effect seems to vary within participants (Roth et al., 2024b). Similarly, explicit spatial-numerical activities such as finger counting vary between individuals. Even in predominantly LR cultures (reading direction and general spatial-directional activities beyond reading), a considerable proportion of participants start finger counting with their right hand (see Hohol et al., 2018 for an overview). To conclude, said effects are typically observed at the group level. We expect Iranians to be more consistent in their RL preferences and Germans in their LR preferences compared to other samples. Based on our cultural directionality hypothesis, Turkish culture includes reverse horizontal directionalities (LR and RL) more than typical LR and typical RL cultures, which could eventually be an explanation for the weaker average SNARC effect in the Turkish sample.

Cultural directional preferences questionnaire (CDPQ)

Directional preferences in different cultures have been widely investigated. The following preferences have been demonstrated to be consistent with reading direction (see Nuerk et al., 2015, for a review): drawing a line LR or RL (Lieblich et al., 1975; Rinaldi et al., 2014), imagining/drawing objects rightward (in RL cultures) or leftward (in LR cultures) (Vaid, 1995), aligning a timeline such that the past is on one side and the future is on the other (Tversky et al., 1991), and preferences regarding the direction/side of action/placement of the agent (Maass & Russo, 2003). Even though previous studies examined all these preferences in isolation, cultural directional preferences have yet to be combined to an assessment. For the present study, we have developed the CDPQ in which we assess horizontal directional preferences across cultures by asking the preferences of participants on certain directionalities. In this way, we could support our initial assumption that the cultural directional preferences in the Turkish sample tend towards RL more compared to the German sample and towards LR more compared to the Iranian sample. In this study, we limit ourselves to horizontal preferences as in the cultures we look at they may directly interfere with the reading direction. Even though other vertical and sagittal cultural directionalities and SNAs have been identified (see Hung et al., 2008; Ito & Hatta, 2004; Winter et al., 2015), they are orthogonal to reading direction in cultures we investigate.

SNARC effect

To examine directional SNAs across three cultures, we used the SNARC effect which is the most widely reported directional SNA. Typically, the SNARC effect is investigated with parity judgment (PJ) and magnitude classification (MC) tasks (see Wood et al., 2008). Although the SNARC effect is observed in both task setups, it is important to differentiate them. Mainly, in PJ, participants are asked to classify numbers as odd or even and, in MC, as small or large. Magnitude processing is a part of the task instruction in MC but is task-irrelevant in PJ. Consequently, the SNARC effect observed in PJ is considered evidence for automatic magnitude processing (see Roth et al., 2024c). In the present study, we examined the SNARC effect across three cultures by using both PJ and MC tasks to reveal whether the cultural directionality hypothesis is applicable only when the magnitude is task-relevant (MC), or whether also applicable when the spatial representation of magnitude is not relevant, and automatically activated (PJ). The latter case would emphasize the strength of the influence of cultural directionalities on the SNARC effect independent from the task.

Concerning the strength of the SNARC effect in these tasks, the findings remain inconsistent. Some studies reported that the SNARC effect is more prominent in MC than in PJ (Cheung et al., 2015; Bae et al., 2009; Fitousi et al., 2009; van Dijck et al., 2009), others reported the reverse pattern; the SNARC effect is stronger in PJ than in MC (Georges et al., 2017; Gevers et al., 2010; Ito & Hatta, 2004) and Didino et al. (2019) reported no difference on the strength of the SNARC effect between the tasks. If the strength of the SNARC effect is related to the mean reaction times (longer RTs related to stronger SNARC; Cipora et al., 2019a, Supplementary Material; Didino et al., 2023; Gevers et al., 2010), we would expect the SNARC effect to be stronger in PJ than MC as RTs in the latter are shorter (Gevers et al., 2010; Wood et al., 2008). However, Wood et al. also reported that the strength of the SNARC effect cannot be only explained by the duration of responses but also by task demands (i.e., whether numbers need to be semantically processed in the task or not). We expected task-relevance of number magnitude (i.e., in MC but not in PJ) to more strongly influence the strength of the SNARC effect than overall RTs (i.e., faster responses in MC than in PJ) so that we hypothesized a stronger SNARC effect in MC than in PJ.

MARC effect

Apart from number magnitude, number parity can also be mapped onto space. More specifically the *Linguistic Markedness of Response Codes* (MARC) effect (Nuerk et al., 2004), describes faster left/right responses to odd/even numbers, respectively. The activation of the parity of a number is not as automatic as the magnitude of the number, though, and therefore, the MARC effect typically only arises in PJ but not in MC.

Linguistic markedness refers to the characteristic of items in frequently used pairs (e.g., good-bad, right-wrong, even–odd, and left–right) that are typically associated with negative and positive polarities (see also Proctor & Cho, 2006). The marked member of the pair is typically more negative and vice versa (e.g., good-unmarked, bad-marked). Nuerk et al. (2004) suggested that the MARC effect is a congruity effect between the markedness of the parity of the number and the response hand (i.e., faster responses for congruent than incongruent situations). Since the MARC effect is influenced by linguistic experiences (e.g., in deaf German signers; Iversen et al., 2004, 2006) and individual differences (e.g., handedness; Huber et al., 2015), one might expect influences of culture as well. To the best of our knowledge, no MARC differences across different reading directions/cultures have been reported. Therefore, in the current study, we examined the MARC effect at the group level in PJ across all samples to examine the cultural influences on the MARC effect.

Finger counting habits

Finger-counting habits seem to be related to the SNARC effect (Fischer & Brugger, 2011). Fischer (2008) reported that participants who count their fingers from the left hand, following an LR direction show a significant SNARC effect, whereas the ones who count from the opposite direction do not. Note that despite several investigations, the effect of finger counting on the SNARC effect is largely inconclusive (e.g., Fabbri, 2013; Hohol et al., 2022; Prete & Tommasi, 2020; Wasner et al., 2014). Lindemann et al. (2011) suggested that the finger-counting direction is consistent with the reading direction of cultures (as many cultural directional preferences). However, this observation has been questioned in later studies showing great within-culture variability in finger counting routines, and in several Western European countries, the majority of participants started from the right hand (e.g., French, Belgian, or Italian participants, see Hohol et al., 2018, for a brief overview). From these aspects, the influence of finger counting on the SNARC effect does not serve as an informative variable, though it could be an explanatory factor in the cultural influences on SNAs. Therefore, we explored finger counting habits and their influence on the SNARC effect across three cultures.

Preregistered hypotheses

The hypotheses were preregistered at https://osf.io/pg5tn. First, we expected directional preferences in the CDPQ to be more LR in German compared to Turkish participants. Furthermore, a stronger RL preference was expected in Iranian compared to Turkish participants. Evidence for these hypotheses would support the cultural directionality hypothesis and our assumption that Turkish is a mixed culture in terms of directionalities (i.e., directional preferences and reading direction).

Furthermore, we expected a regular SNARC effect in German participants based on previous findings for both PJ and MC tasks (see Wood et al., 2008). On the other hand, we had no specific hypotheses for the group-level SNARC effect in Turkish and Iranian samples due to inconclusive or insufficient previous evidence.

Most importantly, we hypothesized a more prominent regular SNARC effect in the German compared to the Turkish sample. Evidence for this would suggest that cultural directional preferences influence SNAs beyond reading direction (which is from LR in both samples). On the other hand, based on the reading direction account, we hypothesized a more prominent SNARC effect in the Turkish compared to the Iranian sample. Evidence for this hypothesis

would reveal an effect of reading (LR vs. RL) by reducing the involvement of cultural directional preferences, which we expect to be RL in both cultures. Therefore, we tested the German < Turkish < Iranian order for SNARC slopes.

We expected the SNARC effects to be more pronounced in the MC compared to PJ because the magnitude processing is task-relevant in MC, unlike PJ.

Method

Participants

A total of 1,144 individuals clicked on the study link. Only 476 of them (157 from Germany, 147 from Turkey, and 172 from Iran) started the experimental tasks and 412 (146 from Germany, 137 from Turkey, 129 from Iran) completed all the experimental tasks.

Participants were university students and recruited via university round mail and social media. Their participation was compensated with €5 or the approximate equivalent in Turkish and Iranian currency. The study was approved by the ethics committee of the Eberhard Karls University of Tuebingen's Department of Psychology.

We applied the following exclusion criteria before the analysis. Unless otherwise noted, all the exclusion criteria below were preregistered. First, we excluded the datasets of 30 participants (5 from Turkey, and 25 from Iran) who reported not wanting to participate seriously but only out of curiosity (seriousness check; e.g., Reips, 2009).

Subsequently, the datasets of 5 participants from Germany who reported not being native German speakers were excluded. Also, there were some bilingual participants in the sample. Since our main hypotheses focus on the culture and reading direction of individuals, we excluded datasets from bilingual participants whose second language is written in reverse direction from their first (LR for Iran and RL for Germany and Turkey), or whose second language belongs to the Middle Eastern (for Germany) or Western culture (for Turkey and Iran). Based on this, the datasets were excluded from 14 bilingual participants (4 from Germany, 3 from Turkey, and 7 from Iran; see Table 1 in the Supplementary Material for detailed information about bilingual participants). This criterion was not preregistered. For the German, Turkish, and Iranian samples, we aimed to recruit participants who were residents of the respective countries. In the Iranian sample, data sets from two participants were excluded, one who reported living in the USA at the time of the experiment and one who reportedly spent the majority of their life in Turkey. These exclusion criteria were also not specified in the preregistration. Since the exclusion criteria for bilingual participants and location were not specified in the preregistration, we also provided the main findings by including datasets from these participants in the Supplementary Materials (Tables 9–14). There was no considerable change in the findings.

Also the datasets of 21 participants (2 from Germany, 7 from Turkey, and 12 from Iran) who described their environment as very/extremely noisy or reported multiple major distractions during the experiments were excluded. Finally, datasets of 23 participants (5 from Germany, 10 from Turkey, and 8 from Iran) were excluded due to having less than 75% valid trials remaining or leaving at least one empty experimental cell to calculate the SNARC effect (number magnitude * response side) in any of the tasks (see Data Preparation section below for further details).

During the preregistration phase, we planned to collect valid data from 130 participants from each sample. However, after applying all the exclusion criteria, the analysis was conducted with 130 German participants (90 women, 37 men, 3 unspecified; 119 right-handed, 8 left-handed, and 3 ambidextrous), 112 Turkish participants (71 women, 40 men, 1 other; 95 right-handed, 12 left-handed, and 5 ambidextrous), and 75 Iranian participants (15 women, 60 men; 63 right-handed, 11 left-handed). Six of the participants did not report their age. All of the remaining participants reported being at least 18 years old. The average age was 24.22 (range = 18—38 years, $SD = 3.69$) for German, 21.75 (range = 18—29 years, $SD = 2.12$) for Turkish, and 24.45 (range = 18—52 years, $SD = 4.29$) for Iranian participants.

Only three participants in the German and one in the Turkish sample reported being able to communicate in a foreign language with RL direction. Forty-six Iranian participants reported being able to communicate in a foreign language with LR direction. See Table 2 in the Supplementary Material for detailed information about the foreign languages that the participants speak. None of the participants' datasets were excluded based on their foreign language.

Materials and procedure

This study was preregistered on the Open Science Framework (OSF: https://doi.org/10.17605/OSF.IO/PG5TN) and administered in the native language of participants (i.e., German, Turkish, and Farsi). The data was collected online between April 11, 2023, and July 31, 2023, simultaneously in all countries. While there are technical and methodological issues with reliability in individual measurement of response times based on operating systems and web browsers used (e.g. Bridges et al., 2020; Garaizar & Reips, 2019), such variation in response times cannot systematically alter support for or against our hypotheses in the present study, as it follows an experimental design with randomization and technical variation over the Internet rather increases external validity compared to lab studies with a reduced set of instrumentation (Reips, 2002, 2021).

Participation was only possible via a desktop computer or laptop. The experimental procedure was implemented via JavaScript (88.1%), CSS (7.9%), HTML (3.3%), and PHP (0.7%) and administered online. An English demo version of the experimental procedure is available at https://osf.io/dk4we. At the beginning of the experiment, participants were asked to participate only if they were at least 18 years old. Furthermore, they were asked to perform the experiment in a quiet room and were informed that they could withdraw at any point by closing the browser tab.

After giving their informed consent, participants were first asked to indicate whether they wanted to seriously participate in the study. After that, there were some demographic questions about their gender, age, native and foreign languages, whether they are actively using their foreign languages, the weekly duration (hours/minutes) of the foreign language use in handwritten, typed, or spoken, country of residence, the country that they spent the majority of their life, familiarity with Western/Middle Eastern cultures, and their handedness. Next, they performed PJ and MC tasks. After the experimental tasks, they responded to the CDPQ, how they represent good and bad on the horizontal space, and finally, to the finger-counting assessment questions. These three parts were only assessed after the completion of the PJ and MC tasks, to avoid triggering a potentially associated spatial mapping. At the end of the study, participants were asked to report on the quietness of the environment and on any major distractions during the participation. English versions of all materials are available at https://osf.io/2w7jp.

PJ and MC tasks

In both PJ and MC tasks, the stimuli were eight numbers from 1 to 9 except 5. Arabic numerals (i.e., 1, 2, 3, 4, 6, 7, 8, and 9) were used for the German and Turkish experiments, whereas Eastern Arabic numerals (i.e., ۱, ۲, ۳, ۴, ۶, ۷, ۸, and ۹) were used for the Farsi experiment. Stimuli were black (0, 0, 0, in RGB notation) on a gray (150, 150, 150, in RGB notation) background. Each trial started with a black square (extended ASCII 254, side length of 72px) for 300 ms as a fixation point. Then the number (Open Sans font, size 72px) was presented until a response was given. Finally, a blank gray screen (150, 150, 150, in RGB notation) for 500 ms concluded the trial.

Each number was presented 40 times in both tasks. Hence, each task consisted of 320 trials divided into two blocks (with a minimum 30-s break and a switch of response-to-key assignment in between) in which each number was presented 20 times in a randomized order. Participants responded to the parity status of the numbers during the PJ task (i.e., whether odd or even) and the magnitude of the numbers during the MC task (i.e., whether smaller or larger than 5) in a binary response setup. The response-to-key assignment was reversed between the blocks which resulted in a MARC-compatible (i.e., odd-left and even-right) and a MARC-incompatible (i.e., odd-right and even-left) block in PJ and a SNARC-compatible (i.e., small-left and large-right) and a SNARC-incompatible (i.e., small-right and large-left) block in MC. As a default, the D and K keys were defined as left and right response keys, respectively. These keys were in similar positions on keyboards that are the most common in the respective country (e.g., D is the third letter from the left and K is the fourth letter from the right on the middle row in all keyboards). Participants had the option to change these response keys to any other two keys that met the criteria of being in the same row and at least four keys' width apart on their keyboard, but none took up this option. The order of the PJ and MC tasks was randomized between participants. Further, each participant was randomly assigned one of two different block orders (SNARC-/MARC-compatible first and SNARC-/MARC-incompatible second in both tasks or SNARC-/MARC-incompatible first and SNARC-/MARC-compatible second in both tasks). Each block was preceded by a short practice session of eight trials (each number presented once). During practice sessions, the required response-to-key assignment was given at the bottom of the screen, and participants received feedback on the accuracy of their responses. Practice sessions were followed by instruction re-iterating the response to the key assignment

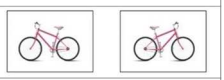
CDPQ Items	
1: object naming "Please write down the name of the objects in the text boxes"	
2: horizontal timeline "Please put the pictures in a meaningful horizontal order"	
3: drawing bike "How would you draw a bike?"	Front leftward or rightward
4: liking bike "Click on the bike you like more"	
5: action moving "Choose the option that depicts '*exchanging gifts*' better."	
6: drawing action moving "How would you draw 'exchanging gifts'?"	LR or RL
7: line drawing "How would you draw a horizontal line?"	LR or RL

Fig. 2 Items in CDPQ. *Bicycle depictions were designed by FreePik

and the start of the experimental block. No feedback and prompts were presented during experimental blocks.

CDPQ

The CDPQ included seven items (see Fig. 2). Item 1 was an object-naming task; participants wrote the names of the horizontally aligned fruit pictures in a vertical (top-down) order. We recorded whether participants followed the LR, RL, or mixed order. In Item 2, participants put three pictures of consecutive events (i.e., an orange being peeled) in a horizontal timeline (Tversky et al., 1991). We recorded whether participants organized the temporal preceding from left to right or right to left or in mixed order. In Items 3 and 4, participants were asked how they would draw a bicycle (leftward or rightward; Vaid, 1995) and to choose the bicycle they liked from two mirror-inverted pictures (one facing left and the other right or indicated that they do not have a preference). In Items 5 and 6, we asked participants how they would draw the scene "exchanging gifts between two people" (the giver on the left and the receiver on the right, vice versa, or no preference) (Maass & Russo, 2003), and to choose the option that depicts this scene better out of two mirror-inverted pictures (action moving from LR or RL). In Item 7, we asked participants to state whether they would draw a horizontal line from LR, RL, or did not have a preference (Lieblich et al., 1975). The presentation order of items was fixed among participants except that the order of Items 3–4 and 5–6 were switched randomly across participants. During the administration of most CDPQ items, participants first saw the instructional text, and only on the next page, they saw the pictures and respond (Items 1, 2, 4, and 5). This specific procedure was preferred to prevent the directional activity of instruction reading from inducing the preference. For items that did not include any pictures but instead only text (Items 3, 6, and 7), participants saw the question and answer options together and responded on the same page. The instructional text and pictures (for items with a picture) and questions and options (for items without a picture) remained on the screen until the participant responded. An English version of CDPQ is available at https://www.numcog.psychologie.uni-tuebingen.de/?demo&en&p=cdq_1.

Apart from the CDPQ, we also asked participants how they would associate the good and the bad with the horizontal space (see Supplementary Material for the findings on this question).

Finger counting assessment

The participants' finger-counting behavior was measured with two questions. First, "How do you count your fingers from 1 to 10?" for which the responses were "I prefer to start counting with the fingers of my right hand", "I prefer to start counting with the fingers of my left hand", "I do not know or I do not have any preferred hand", or "I prefer not to answer". The second question was "How stable are your finger-counting habits?" and the responses were "I always start counting with this hand", "I usually start counting with this hand", "I do not know or I do not have a preferred hand or stable habit", or "I prefer not to answer".

Analyses

In addition to all the frequentist tests described below, we report corresponding Bayes Factors to differentiate evidence of the absence of an effect from the absence of evidence for an effect. To calculate Bayes Factors, we used the default r-scale of 0.707 as an uninformed prior in the Cauchy distribution. To interpret Bayes Factors, we follow Dienes' (2021) recommendations, namely, a BF_{10} above 3 or 10 will be treated as moderate or strong evidence for the alternative hypothesis, and a BF_{10} below 1/3 or 1/10, as moderate or strong evidence for the null hypothesis.

PJ and MC tasks

Data preparation All steps of the data preparation followed our preregistered protocol. Only experimental trials were analyzed. First, incorrect trials were excluded (4.92% in PJ and 3.67% in MC). Next, RTs shorter than 200 ms and longer than 1500 ms were discarded (1.69% in PJ and 1.13% in MC). Later, within each participant, a sequential filtering approach (Cipora & Nuerk, 2013; Cipora et al., 2019a) was applied to trim outlier RTs in the data separately for PJ and MC tasks. Specifically, RTs outside ± 3 SD from the individual's mean RT were excluded. This filtering procedure was repeated until means and SDs no longer changed, resulting in an exclusion of 4.15% of the data in PJ and 4.97% of the data in MC. For 23 participants (5 from the German, 10 from the Turkish, and 8 from the Iranian sample), more than 25% of the dataset was discarded in at least one of the tasks, and these participants' data were excluded from the analysis. This left data from 130 German, 112 Turkish, and 75 Iranian participants for further analysis. After RT preprocessing, 89.24% of this data in PJ and 90.23% of their data in MC were retained.

Data analyses All the data analyses described here were performed by following the preregistration unless otherwise stated. In all analyses, a significance level of $\alpha = 0.05$ was used. We had not preregistered how to correct α for multiple tests. Bonferroni-Holm method was applied to correct for multiple tests in all analyses unless noted otherwise.

We investigated the SNARC effect in both PJ and MC tasks, and the MARC effect only in the PJ task, by using a repeated-measures regression approach (proposed by Lorch

& Myers, 1990; adapted to the SNARC effect by Fias et al., 1996). First, dRTs (mean right-hand response RT minus mean left-hand response RT) were calculated for each number separately within each task for each participant.

Based on the continuous shape of the SNARC effect in PJ (Wood et al., 2008), a linear predictor was defined with the values 1, 2, 3, 4, 6, 7, 8, and 9. Furthermore, the parity of numbers was added to the regression model of the PJ task as a contrast-coded predictor (with -0.5 for odd and 0.5 for even numbers) to investigate the MARC effect (see Nuerk et al., 2004, 2005; Cipora et al., 2019a). Number parity is orthogonal to the number magnitude in the given stimuli set, therefore, including the parity as an additional predictor in the regression model does not affect the parameter estimate of number magnitude (Cipora et al., 2019a).

For the MC task, average dRTs were calculated for small (1, 2, 3, and 4) and large (6, 7, 8, and 9) numbers. Based on the categorical shape of the SNARC effect in MC tasks (Wood et al., 2008), a categorical predictor was defined with the values -0.5 for small numbers and + 0.5 for large numbers.

The resulting coefficients for number magnitude from the individual analyses were considered as unstandardized SNARC slopes. In the PJ task, negative unstandardized SNARC slopes reflect the increase in RT advantage (in milliseconds) of the right hand over the left per increase in number magnitude in one unit (because number magnitude is a linear predictor). Coefficients for the number parity predictor were considered as unstandardized MARC slopes. Negative unstandardized MARC slopes reflect the increase of the RT advantage of the right hand over the left hand for even numbers compared to odd numbers. In the MC task, negative unstandardized SNARC slopes reflect the increase of the RT advantage of the right hand over the left hand for large numbers over small numbers.

In addition, standardized SNARC and MARC slopes (Fisher-z-transformed standardized regression weights) were calculated (see Hochmann et al., 2024; Cipora et al., 2019a). Standardized slopes can be considered as the individual effect sizes of the SNARC and MARC effects and are not influenced by individual reaction time or individual variance within reaction time (Cipora et al., 2019b). More negative slopes for both magnitude and parity predictors correspond to the stronger SNARC and MARC effects, respectively.

Unstandardized and standardized SNARC and MARC slopes were compared against zero using two-sided one-sample *t*-tests to examine the group-level SNARC and MARC effects. The task order (PJ first and MC first) and block order (SNARC-/MARC-compatible first and SNARC-/MARC-incompatible first) effects were examined using independent-sample *t*-tests in each sample (see Supplementary Material Tables 4 and 5 for the findings of this analysis).

Subsequently, unstandardized SNARC slopes were compared across samples with a Jonckheere-Terpstra test for monotone trend with 2000 permutations, one-sided for the German < Turkish < Iranian alternative hypothesis. This non-parametric test, in comparison to linear contrast in ANOVA, only assumes a monotone trend in the data, while it does not assume equidistance between group means. Therefore, it best reflects our hypothesis as we could not assume that the German vs. Turkish mean difference would be the same as the Turkish vs. Iranian difference.

This was followed by comparison of the PJ and MC in terms of the size of the unstandardized SNARC effect. It is not possible to compare a categorical and a linear slope because the former is naturally much larger when considering the same underlying dataset, therefore, additional individual regression analyses were performed for the MC task with magnitude being a linear predictor just as for the PJ task. Then the unstandardized SNARC slopes from the continuous models were compared with paired samples *t*-tests within each sample. Moreover, to reveal a possible relationship between the SNARC effects observed in the two tasks, the correlation of the unstandardized SNARC slopes between the tasks (continuous for parity judgment and categorical for magnitude classification) was examined across samples.

Furthermore, H0 bootstrapping analysis (Cipora et al., 2019b; see also Roth et al., 2024b) was performed to calculate the individual prevalence of SNARC and MARC effects within each sample. The analysis calculates the probability of finding the SNARC/MARC slopes as we observed during the experiments for each participant, assuming there was no association between magnitude/parity of the number and response side (i.e., if the null hypothesis was true). Separately for the tasks, within each individual's dataset, two sets of 20 responses were randomly sampled with replacement from all RTs to a given number. One of the sets was considered to be left-hand responses and the other, right-hand responses. Next, unstandardized SNARC and MARC slopes were calculated using these sampled data. This procedure was repeated 5000 times, resulting in a bootstrapped slope distribution for each participant. Next, empirically obtained slopes were examined to see whether they were outside or inside of the mid-90% of bootstrapped slope distributions (i.e., 90% H0 CIs). Based on these CIs, the individual slopes were classified as reliable SNARC or MARC effect (empirically obtained slope < lower bound of H0 CI), reliable reverse SNARC or MARC effect (empirically obtained slope > upper bound of H0 CI), or unreliable SNARC or MARC effect (empirically obtained slope within H0 CIs; see Hohol et al., 2020, 2022; van Dijck et al., 2022). A χ^2 test was calculated to investigate whether the proportions differ between samples.

We examined the impact of foreign languages (RL for German and Turkish and LR for Farsi) on the SNARC effect by correlating the duration (in minutes) of the foreign language use (handwritten, typed, or spoken) and the unstandardized SNARC slopes from the two tasks (see Supplementary Material for the findings of this analysis).

CDPQ

For each item, a 3 (sample: German, Turkish, and Iranian) × 3 (directionality: LR, NP (no preference), and RL) χ^2analysis was performed. Fisher's Exact Test was performed instead of a χ^2 when expected frequencies in any of the cells were lower than 5. In addition to the descriptive analyses, a sum score was calculated for each participant for the CDPQ. First, each LR response was scored as -1, and each RL response as 1. If there was no LR or RL order in the first two items (i.e., object naming task and timeline), these responses were scored as 0. In Items 3–7, "I have no clear preference" responses were also scored as 0. Therefore, an individual sum score across the seven items could vary between -7 (very strong LR preference) and 7 (very strong RL preference). The participants' sum score was compared across samples with a one-sided Jonckheere-Terpstra test for monotone trend with 2000 permutations for the alternative hypothesis German < Turkish < Iranian. Finally, the correlation between the unstandardized SNARC slopes and the sum scores of the CDPQ was examined across samples.

Finger counting

Independent samples *t*-tests were performed in each sample to examine the possible effect of finger counting direction (left- vs. right-starters) on the unstandardized SNARC slopes. Furthermore, a Jonckheere-Terpstra test for monotone trend with 2000 permutations was performed separately for left- and right-starters in each sample to reveal whether finger counting stability has any influence on the unstandardized SNARC slopes (see Supplementary Material Tables 7 and 8 for the findings of this analysis).

Results

The SNARC effect and the MARC effect

The results of all the *t*-tests performed to investigate the presence of the SNARC and the MARC effect across all samples in the PJ task are shown in Table 1. There was a significant SNARC effect at the group level across all samples (Fig. 3). Note that the Bayesian *t*-test showed inconclusive evidence regarding both the unstandardized and the standardized SNARC effect in the Iranian sample. The distribution of the unstandardized SNARC slopes for each sample can be seen in Fig. 4. Interestingly, no significant unstandardized or standardized MARC effects were observed in any of the samples.

The results of the *t*-tests performed to investigate the presence of the SNARC effect across all samples in the MC task are shown in Table 2. There was a significant SNARC effect at the group level in German and Turkish samples with unstandardized and standardized measures, but no significant unstandardized or standardized SNARC effects in the Iranian sample. Furthermore, Bayesian *t*-tests showed inconclusive evidence for the alternative hypothesis of the

Table 1 The SNARC effect and the MARC effect in PJ

Measure	Sample	Mean Slope (SD)	*t*	*p*	BF_{10}
SNARC	German	−6.03 (6.68)	−10.30	< 0.001*	3.33e + 15*
	Turkish	−3.43 (8.63)	−4.21	< 0.001*	319.74*
	Iranian	−1.69 (6.72)	−2.18	0.033*	1.16
Standardized SNARC	German	−0.43 (0.45)	−10.72	< 0.001*	3.50e + 16*
	Turkish	−0.21 (0.46)	−4.79	< 0.001*	2773.61*
	Iranian	−0.11 (0.48)	−2.03	0.045*	0.89
MARC	German	−8.39 (68.36)	−1.40	0.164	0.25*
	Turkish	3.71 (83.44)	0.47	0.639	0.12*
	Iranian	−4.92 (72.57)	−0.59	0.559	0.15*
Standardized MARC	German	−0.14 (0.83)	−1.90	0.060	0.56
	Turkish	0.04 (0.84)	0.45	0.656	0.12*
	Iranian	−0.10 (0.81)	−1.06	0.293	0.22*

All tests were two-sided and against zero

df was 129 for the German, 111 for the Turkish, and 74 for the Iranian sample

* $p < 0.05$ indicating a significant finding (frequentist), $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis, and $BF_{10} > 3$ indicating a conclusive finding for the alternative hypothesis (Bayesian)

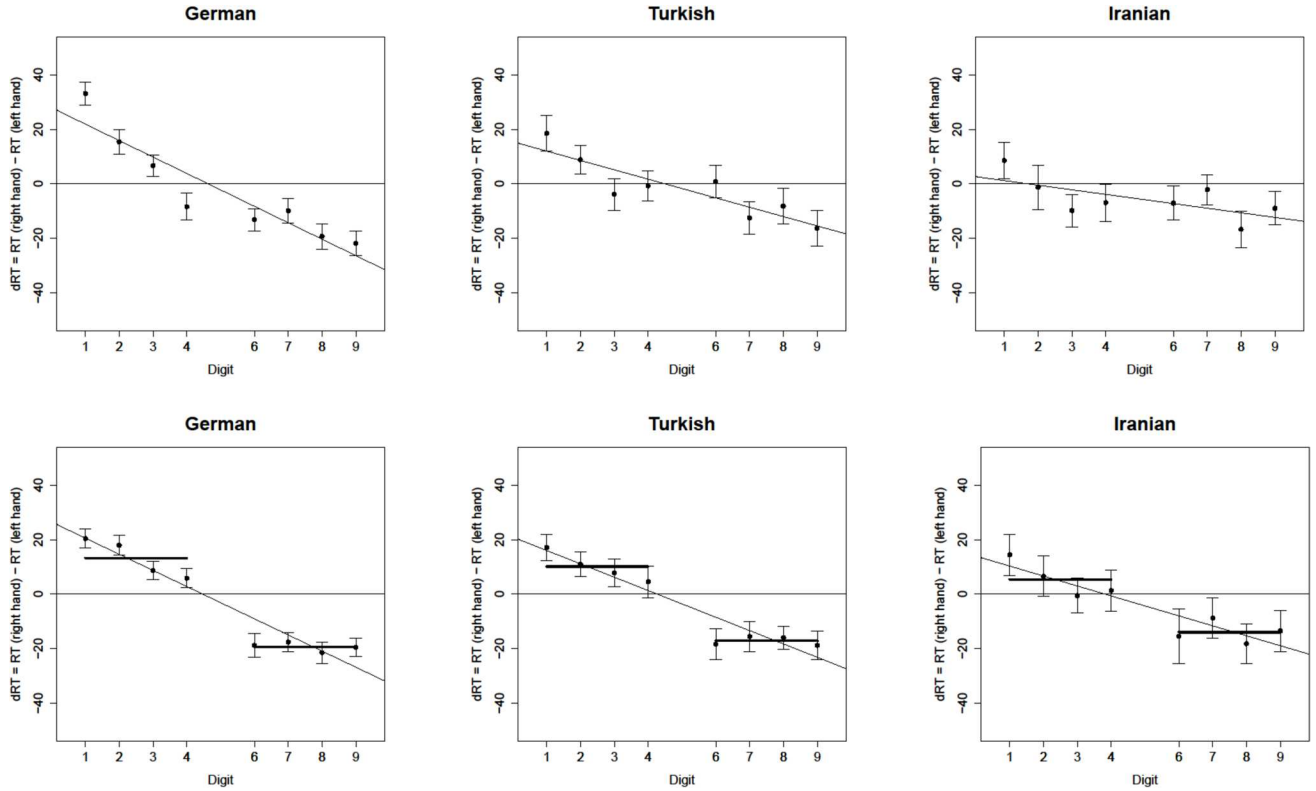


Fig. 3 The SNARC slopes in PJ (upper panel) and MC (lower panel) across three samples. The SNARC effect is characterized by a negative slope, indicating larger digits were responded to faster with the right than with the left-hand

presence of the unstandardized and standardized SNARC effect in the Iranian sample.

Comparing the SNARC effect across samples

As expected, the one-sided Jonckheere-Terpstra test (2000 permutations) revealed a significant monotone trend of German < Turkish < Iranian SNARC slopes in PJ, $T_{JT} = 19{,}905$, $p < 0.001$. One-sided independent samples Welch *t*-test showed a significant difference for the SNARC slopes of German and Turkish samples ($t(207.48) = -2.59$, $p = 0.005$) but not for the Turkish and Iranian samples ($t(180.75) = -1.55$, $p = 0.061$). Comparing the SNARC slopes with independent Bayesian post-hoc *t*-tests provided moderate evidence for a stronger SNARC effect in the German than in the Turkish sample ($BF_{10} = 3.68$), but evidence in the Turkish-Farsi comparison remained inconclusive ($BF_{10} = 0.44$).

Similarly, the one-sided Jonckheere-Terpstra revealed a significant result for the directional alternative hypothesis of German < Turkish < Farsi SNARC slopes in MC, $T_{JT} = 17{,}835$, $p = 0.041$. One-sided independent samples Welch *t*-test showed no significant difference between German-Turkish ($t(220.26) = -0.60$, $p = 0.276$) and Turkish-Farsi ($t(122.27) = -0.56$, $p = 0.288$) comparison on the SNARC slopes. Comparing the SNARC slopes with independent Bayesian *t*-tests revealed moderate evidence for the null hypothesis in the German-Turkish comparison ($BF_{10} = 0.17$) and Turkish-Farsi comparison ($BF_{10} = 0.19$).

Comparing PJ and MC tasks

Calculating the SNARC slopes of the MC task with the magnitude being a continuous predictor yielded similar findings to the analysis in which the magnitude was a categorical predictor. With the continuous predictor, for the German and Turkish samples, there was a significant SNARC effect at the group level (German: $t(129) = -6.36$, $p < 0.001$, $M = -5.92$, $SD = 10.61$; Turkish: $t(111) = -4.22$, $p < 0.001$, $M = -4.90$, $SD = 12.30$) but not for the Iranian group ($t(74) = -1.81$, $p = 0.074$, $M = -3.65$, $SD = 17.46$). Note that categorical predictors typically fit the MC task data better than continuous (Wood et al., 2008). In the present findings for German and Turkish data, the categorical predictor (adj. R^2 was 0.92 and 0.93, respectively) was slightly a better fit than the continuous (adj. R^2 was 0.91 and 0.90, respectively) in MC. Interestingly though, continuous (adj. $R^2 = 0.83$) was a better fit than categorical (adj. $R^2 = 0.77$) for Iranian data in MC. As can be seen in the upper panel of Fig. 3, this finding was due

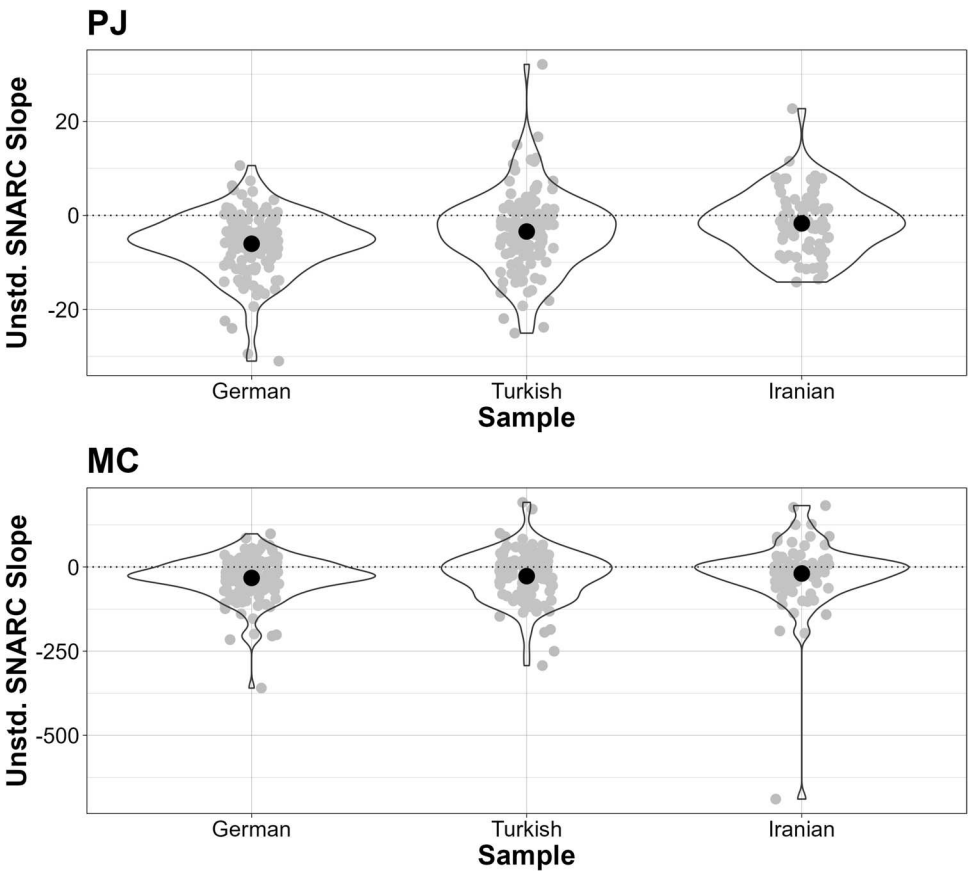


Fig. 4 The distribution of the unstandardized SNARC slopes in each sample

Table 2 The SNARC effect in MC

Measure	Sample	Mean Slope (SD)	*t*	*p*	BF_{10}
SNARC	German	−32.58 (63.38)	−5.86	< 0.001*	299,484.60*
	Turkish	−27.26 (73.84)	−3.91	< 0.001*	113.51*
	Iranian	−19.42 (104.92)	−1.60	0.113	0.43
Standardized SNARC	German	−0.34 (0.61)	−6.31	< 0.001*	3.50e + 16*
	Turkish	−0.23 (0.59)	−4.04	< 0.001*	2773.61*
	Iranian	−0.13 (0.60)	−1.83	0.072	0.89

All tests were two-sided and against zero

df was 129 for the German, 111 for the Turkish, and 74 for the Iranian sample

*$p < 0.05$ indicating a significant finding (frequentist), $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis, and $BF_{10} > 3$ indicating a conclusive finding for the alternative hypothesis (Bayesian)

to the first four numbers, but not the larger numbers, being continuously arranged in dRTs.

The SNARC slopes of the PJ and MC obtained from continuous magnitude predictors were compared separately in each sample, and results show no significant difference between PJ and MC slopes in any of the samples (German: $t(129) = 0.10$, $p = 0.918$, $BF_{10} = 0.10$; Turkish: $t(111) = -1.00$, $p = 0.318$, $BF_{10} = 0.17$; Iranian: $t(74) = -0.94$, $p = 0.353$, $BF_{10} = 0.19$).

Correlation between the PJ and MC SNARC slopes

There was no significant correlation between the PJ and MC unstandardized SNARC slopes in any of the samples (German: $r = -0.02$, $p = 0.822$, $BF_{10} = 0.21$; Turkish: $r = -0.08$, $p = 0.386$, $BF_{10} = 0.31$; Iranian: $r = 0.10$, $p = 0.402$, $BF_{10} = 0.37$).

H0 bootstrap analysis

Table 3 summarizes the individual prevalence of the SNARC effect in all samples for both tasks. For the unstandardized SNARC effect in PJ, in the German sample, the majority showed a reliable regular SNARC effect, unlike in the Turkish and Iranian samples. A higher number of participants showed an unreliable regular or reversed SNARC effect than those who showed a reliable SNARC effect. A χ^2 analysis confirmed a significant association between the culture and the number of participants who showed a reliable SNARC effect, $\chi^2(4) = 18.67$, $p < 0.001$.

Table 3 Individual prevalence of the reliable SNARC effect calculated with H0 bootstrapping in PJ and MC

Task	Sample (N)	Reliable SNARC (*N*)	Reliable reverse SNARC (*N*)	Unreliable SNARC (*N*)
PJ	German (130)	56.92% (74)	4.62% (6)	38.46% (50)
	Turkish (112)	37.50% (42)	11.61% (13)	50.89% (57)
	Iranian (75)	29.33% (22)	12.00% (9)	58.67% (44)
MC	German (130)	56.92% (74)	16.15% (21)	26.92% (35)
	Turkish (112)	46.43% (52)	19.64% (22)	33.93% (38)
	Iranian (75)	38.67% (29)	21.33% (16)	40.00% (30)

Table 4 Individual prevalence of the reliable MARC effect calculated with H0 bootstrapping in PJ

Sample (*N*)	Reliable MARC (*N*)	Reliable reverse MARC (*N*)	Unreliable MARC (*N*)
German (130)	41.54% (54)	23.85% (31)	34.62% (45)
Turkish (112)	35.71% (40)	37.50% (42)	26.79% (30)
Iranian (75)	36.00% (27)	22.67% (17)	41.33% (31)

Table 5 Two-sided independent-sample t-test findings comparing the unstandardized SNARC slope of left- and right-starters

Sample	Task	*t*	*df*	*p*	BF_{10}
German	PJ	0.87	119.88	0.389	0.27*
	MC	−1.08	117.09	0.283	0.33*
Turkish	PJ	−0.46	70.07	0.644	0.24*
	MC	0.25	86.67	0.801	0.22*
Iranian	PJ	−0.20	33.71	0.841	0.28*
	MC	0.59	36.50	0.562	0.31*

* $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis

For the unstandardized SNARC effect in MC, the number of participants in the German and Turkish samples who showed a reliable regular SNARC effect was higher than those who showed an unreliable regular or reversed SNARC effect. On the other hand, in the Iranian sample, there was a more even balance between those showing unreliable and reliable regular SNARC effects. The χ^2 analysis revealed no significant association between the culture and the number of participants who showed reliable SNARC effect in the MC task, $\chi^2(4) = 6.86$, $p = 0.144$.

Table 4 summarizes the individual prevalence of the MARC effect in all samples for the PJ task. The reliable regular MARC effect was more prominent in the German sample compared to the other samples. The χ^2 analysis revealed no significant association between the sample and the number of participants showing a reliable MARC effect in the PJ task, $\chi^2(4) = 8.67$, $p = 0.070$.

Finger counting habits

In the German and Iranian samples, most participants reported counting their fingers starting from the right hand; interestingly, the majority of the participants in the Turkish sample reported starting from the left hand (see Supplementary Material Table 6). Most of the participants reported that they are always or usually stable in their finger-counting habits (see Supplementary Material Table 7). Unstandardized SNARC slopes of PJ and MC did not differ between left- and right-starters in any of the samples, with moderate Bayesian evidence indicating that finger-counting direction had no influence on the SNARC effect (see Table 5).

Furthermore, a series of two-sided Jonckheere-Terpstra tests showed no difference in unstandardized SNARC slopes based on the finger counting stability for either left- or right-starters (see Supplementary Material Table 8).

CDPQ

Figure 5 summarizes the directional preferences in CDPQ items. χ^2 analysis performed for each item is shown in Table 6. Interestingly, samples did not differ based on liking a leftward or rightward-facing bicycle, although they showed differences in their preferences for drawing a bike. Post-hoc comparisons of samples for each item can be found in Tables 7 and 8.

CDPQ total score

In line with the unstandardized SNARC slopes, the one-sided Jonckheere-Terpstra test (2000 permutations) revealed a significant monotone trend of German < Turkish < Iranian total CDPQ scores, $T_{JT} = 24{,}156.50$, $p < 0.001$, indicating

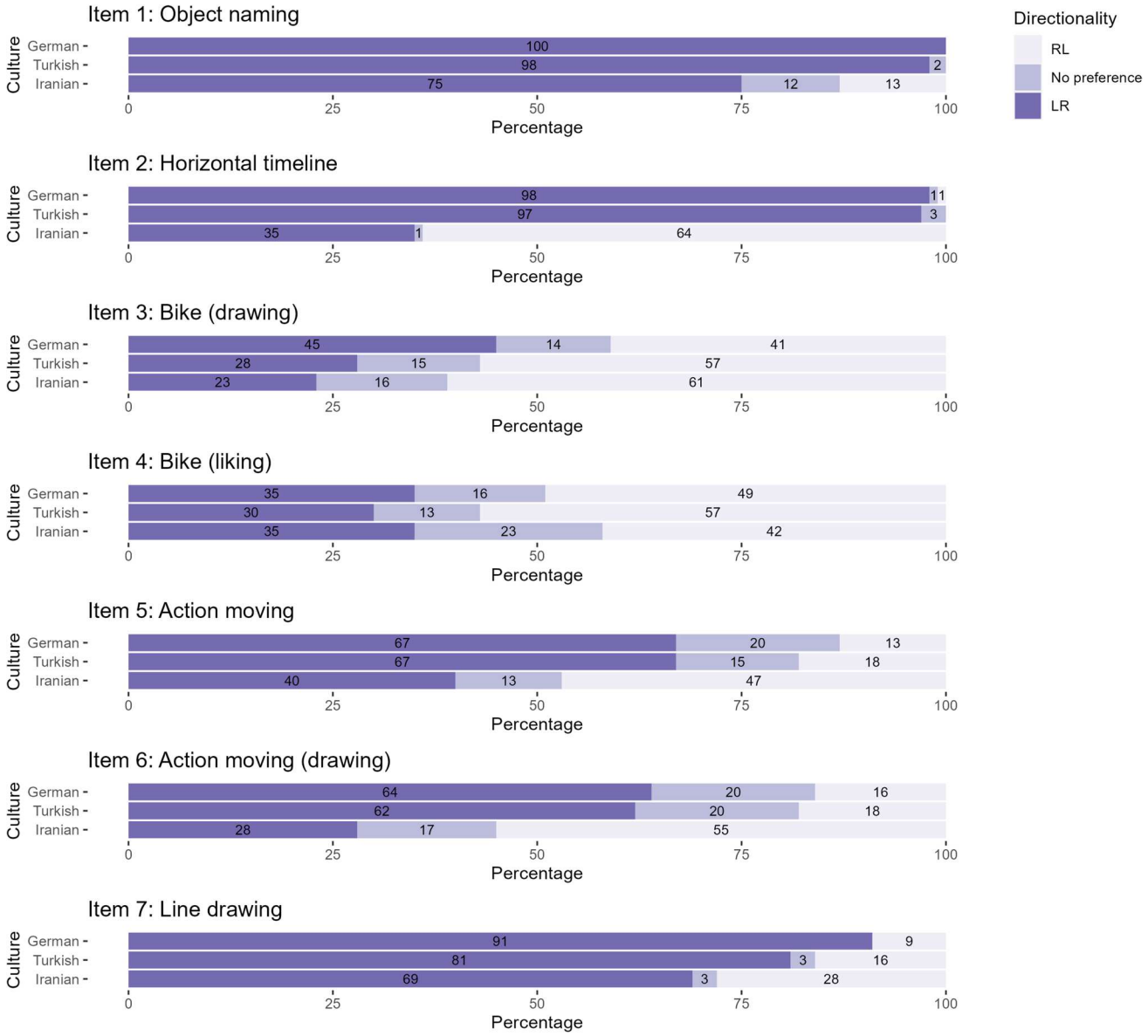


Fig. 5 Directional Preferences in CDPQ Items

that LR directionality was decreasing in a monotone way from German (M = -3.65) through Turkish (M = -2.96) to the Iranian sample (M = 0.07).

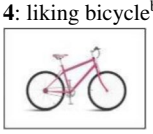
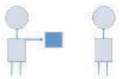
While the between-group differences were significant, the correlations between the participants' unstandardized SNARC slopes and the participants' total CDPQ scores were not significant in any of the tasks or within any of the samples (PJ: German: r = -0.07, p = 0.435, BF_{10} = 0.27; Turkish: r = -0.06, p = 0.549, BF_{10} = 0.26; Iranian: r = -0.06, p = 0.617, BF_{10} = 0.30; MC: German: r = 0.21, p = 0.015, BF_{10} = 3.47, *adj.* p = 0.089; Turkish: r = 0.06, p = 0.542, BF_{10} = 0.26; Iranian: r = -0.10, p = 0.371, BF_{10} = 0.38). However, for Germans in the MC, the BF revealed anecdotal evidence, and frequentist testing would have been significant if tested one-sided (which was not specified as such in the preregistration). This provides a hint that the relationship between cultural directional habits and SNARC slopes may deserve more investigation in the future. On the other hand, given the relatively inconclusive results, we cannot draw strong conclusions from this data.

Discussion

Overview

The present study tested the cultural directionality hypothesis that SNAs are influenced by individuals' socio-cultural

Table 6 X^2 analysis findings for each item in the CDPQ

Items	χ^2	*df*	*p*
1: object naming[a]	–	-	<0.001*
2: horizontal timeline[a]	-	-	<0.001*
3: drawing bicycle	13.31	4	.010*
4: liking bicycle[b] 	4.88	4	.300
5: action moving 	33.80	4	<0.001*
6: action drawing	44.59	4	<0.001*
7: line drawing[a]	-	-	<0.001*

[a]Fisher's Exact test was performed when the expected frequencies of a cell were lower than 5

[b]Bicycle demonstration was designed by FreePik

*p < .05 indicating a significant finding

Table 7 Fisher's Exact Test post-hoc comparisons for CDPQ items

Item	Comparison	p
1: object naming	German- Iranian	<0.001*
	Turkish-Iranian	<0.001*
	German- Turkish	.213
2: horizontal timeline	German- Iranian	<0.001*
	Turkish-Iranian	<0.001*
	German- Turkish	.820
7: drawing a line	German- Iranian	<0.001*
	Turkish-Iranian	.118
	German- Turkish	.044*

Bonferroni correction was applied in post-hoc tests

* *p* < 0.05 indicating a significant finding

environment, including reading direction and many other common cultural directional preferences. To test this hypothesis, we measured the SNARC effect with PJ and MC tasks in three different cultures; a predominantly LR (German), a predominantly RL (Iranian), and a mixed (Turkish) culture regarding cultural directional preferences and reading habits. After following an RL script for almost 600 years, Turkey

Table 8 X^2 analysis significant post-hoc findings

Item	Sample	Directionality	p
3: bike (drawing)	German	LR	<0.01 (higher)*
		RL	<0.01 (lower)*
5: action moving	German	RL	<0.001 (lower)*
	Iranian	RL	<0.001 (higher)*
		LR	<0.001 (lower)*
6: action moving (drawing)	German	RL	<0.001 (lower)*
	Iranian	RL	<0.001 (higher)*
		LR	<0.001 (lower)*

Only cells that are significantly lower or higher than expected frequencies were summarized in the table

Bonferroni correction was applied in post-hoc tests

* *p* < 0.05 indicating a significant finding

changed to an LR script direction as recently as 100 years ago. We hypothesized that, despite this change, some cultural directional preferences prevailed in RL due to six centuries of that tradition. To confirm that cultural directional preferences are indeed more RL in Turkish culture compared to German and more LR compared to Iranian culture, we administered the newly developed CDPQ. Findings supported this categorization, the CDPQ total score being more strongly LR for German, more strongly RL for Iranian, and in between for Turkish participants. Results of the PJ and MC tasks were in line with CDPQ findings, showing that SNARC effects in both tasks were strongest for German, weakest for Iranian, and intermediate for Turkish participants. However, this group-level effect was not reflected in individual-level correlations. Overall, findings showed that directional preferences differed between cultures and the emergence of adult SNAs in two tasks was consistent with those cultural directional preferences.

Cultural directional preferences across countries

As expected, the CDPQ showed that preferences were generally LR in German; however, surprisingly, they were not generally RL in Iran (see Faghihi & Vaid, 2023 for a review), but rather, indicating no preference overall. We could conceive two possible explanations for this.

First, it has been argued that LR directionality is initially rather innate (e.g., Felisatti et al., 2020) and only later overwritten by culture (e.g., McCrink & de Hevia, 2018). It is possible that this overwriting occurs only partially so that responses in the CDPQ are influenced by both innate LR directionality and cultural RL directionality, leading to an actual null effect. Note in this respect that the LR/RL directionality for Iranians differs greatly depending on the focus of attention. While object naming and line drawing were from LR, even in most Iranian participants, the action

moving items, horizontal timeline, and bicycle (drawing) were predominantly from RL. Thus, it is possible that certain activities are more influenced by culture, and others, are less so.

Second, the world culture portrayed in movies and video games, experienced in international airports, computers, and mobile phone systems comes predominantly from LR cultures (e.g., the US). Individuals exposed to it are more likely to be confronted with LR culture in images and experience, which was the subject of our CDPQ. This global LR directionality might influence cultural habits in addition to their own culture. Some cultural aspects might be more RL because they are influenced more by their own culture (e.g., rarely encountered via international media, travel, and technology), while those more frequently encountered in international culture may be more LR. Since global culture and language influence many aspects of national languages (e.g., via loan words) and culture (e.g., via international music), it is clearly conceivable that this holds for directionality as well. It is therefore impossible to distinguish the two hypotheses in urban samples exposed to global culture, and future research should disentangle these by studying remote populations with RL culture and restricted or no access to global culture.

Overall, the preferences of German and Iranian participants for most items were in line with the previous findings of LR and RL cultures, respectively. The horizontal timeline and action moving items generally followed the respective culture direction as demonstrated previously (see Tversky et al., 1991; Mass & Russo, 2003). In the bicycle drawing, German participants reported a greater preference to draw the front-facing leftward compared to other samples. When the front is leftward, the bicycle appears moving RL, which might seem counterintuitive for an LR culture. On the other hand, this finding is in line with previous research with adults and children that demonstrated the tendency to draw the objects so that they appear moving RL in LR cultures, and so that they appear moving LR in RL cultures (Kebbe & Vinter, 2013; Vaid, 1995). The significant difference between samples in bicycle drawing disappeared in the expression of bicycle liking, suggesting that this directional bias difference across samples was specific to drawing habits, but not necessarily generalized to liking for objects. One possible explanation of this discrepancy between drawing and liking on directional preferences is that since people tend to start from the left side of the paper in LR cultures, they might place the most important and detailed part (i.e., the front) on the left side and continuing the drawing (i.e., the back) toward the right side. Correspondingly, we would expect the exact opposite pattern in RL cultures. Note that, in the present study, participants did not actually draw but only reported how they would draw the bicycle, which might underestimate their directional bias compared to lab studies. Interestingly, there was an overall LR preference in the object naming, including for Iranian participants. This finding is interesting since Shaki et al. (2012) showed RL preference in a similar task (i.e., object counting) in RL cultures (see Shaki & Fischer, 2023 for the influence of object counting direction on the attentional SNARC effect). On the other hand, a study by Shaki and Gevers (2011) suggested that cultural characteristics can manifest themselves in reverse directions for magnitude and ordinal information processing. In their study, magnitude information was associated with LR processing but ordinal information was with RL processing in Hebrew participants. In addition to that, although the tasks in the mentioned studies seem superficially similar, object naming in the present study and object counting in Shaki et al. (2012) might include different directionalities: object counting requires active object numbering by pointing at objects, whereas in the present study, participants identified pictures of the objects on a computer screen and wrote the words for each object, an activity which might be influenced by LR computer screen scanning habits on a more global scale. Although significantly less prominent in Iranian participants, there was an overall LR preference in line drawing, in contrast to Lieblich et al.'s (1975) study, and again, this difference could be due to the online setup of the questionnaire.

Our initial expectation was that preferences in Turkish culture would be closer to Iranian rather than German culture, but this did not seem to be the case in the present study. The total scores were negative (i.e., indicating LR direction) for both German and Turkish participants and close to zero (i.e., indicating no preference) for Iranian participants. More specifically, the preferences in object naming, horizontal timeline, and action-moving items were similar to those in Germany. On the other hand, Turkish and Iranian preferences were closer to each other compared to German preferences in bicycle drawing and line drawing. This pattern suggests that certain cultural directional preferences in Turkish culture more tightly linked to the reading direction (i.e., habits that require scanning of horizontally aligned objects/events) tend more to LR. Other preferences that are less influenced by reading direction and more ingrained in the culture may tend towards more RL patterns, because the LR reading direction has only been relatively recently established in Turkey, i.e., for about four generations.

As a result, the CDPQ successfully captured differences in cultural directional preferences across samples in an online setup. This suggests the relative effectiveness of this approach, namely, the pseudo-experimental manipulation of testing of three different cultures, with one being predominantly LR in cultural directional preferences, one being predominantly RL, and the third being mixed between

reading direction and cultural directional preferences. Based on these results, the next question was whether such cultural differences would affect SNAs.

Directional SNAs are influenced by cultural directional preferences

The differences between German and Turkish culture in the SNARC effect, despite the common LR reading direction, strongly suggest that factors other than reading direction are potentially involved in the emergence of SNAs. Turkish CDPQ scores are less LR than German ones, suggesting that these factors could be cultural directional preferences. The influence of cultural directional preferences was previously proposed in several studies (see Nuerk et al., 2015; Patro et al., 2016a, 2016b; Göbel et al., 2018) to explain the emergence of SNAs in preschool children. In the present study, we further suggest that the influence of cultural directional preferences is not merely a transient phase during early preschool development, where reading direction is not (well) developed. Instead, it seems that, even in adults, cultural directional preferences influence SNAs beyond reading direction.

The hypothesized reason for the regular but weaker (compared to German culture) SNARC effect in Turkish culture is that there are RL cultural directional preferences (at least more RL as compared to Germany), even though the reading direction is LR. The underlying reason for the absence of strong and consistent LR directionalities in Turkish culture could be related to the script change around a hundred years ago. However, it could be argued that participants in the present sample were not influenced by this change directly. Note that in 1928 (i.e., just before the time of the Latin alphabet revolution), only 8% of the Turkish population was literate (Yılmaz, 2011), indicating that, even at that time, only a small portion of the population was directly influenced by the change. Importantly though, the reading direction of a culture infiltrates many organizations in the cultural environment and typically covaries with the cultural directionalities. Therefore, (semi-) illiterate individuals can be still influenced by their socio-cultural environment, which naturally involves directionalities (see Nuerk et al., 2015, for a review). Before the Latin alphabet revolution, Turkish culture was probably a more consistent RL culture due to not only the RL Arabic-Farsi alphabet but also the Middle-Eastern and Islamic RL influence on the cultural environment. The literacy rate increased dramatically after the Latin alphabet revolution, eventually influencing the directional preferences in the cultural environment. Nevertheless, some RL habits might have remained, especially those not directly linked to reading direction. Considering that the reading direction is LR, these remaining RL habits might create an inconsistency in associations of space and sequence or space and magnitude (see Patro et al., 2016a, for how such influences may develop), which may be the cause of weaker LR SNAs in Turkish culture.

The differences between Turkish and Iranian culture in the SNARC effect in both tasks suggest that reading direction might also be involved in the emergence of the SNAs. As demonstrated previously in the literature, RL reading habit has a weakening influence on the regular SNARC effect compared to LR reading habit (e.g., Dehaene et al., 1993; Shaki & Fischer, 2008; Lopiccolo & Chang, 2021) and has even been claimed by some to reverse the effect (e.g., Shaki et al., 2009; Zebian, 2005). Nevertheless, when we evaluate the literature overall, the influence of reading habits on the SNARC effect remains unclear due to inconsistent findings among RL readers as well as the absence of evidence of a causal relationship between reading activity and number processing (see Pitt & Casasanto, 2020).

The group-level regular SNARC effect was significant for German and Turkish participants in both PJ and MC tasks. This finding was unsurprising for the German sample since the effect has been demonstrated in LR cultures several times with both tasks (see Wood et al., 2008). In contrast, Bulut et al. (2023) reported no SNARC effect in a Turkish sample with a PJ task¸ however, it should be noted that there are three major differences between Bulut et al. (2023) and the current study. First, in the present study, the number of participants was considerably higher, second, the number of trials per response-to-key assignment was two-fold as in Bulut et al. (2023). These two parameters are highly influential in increasing the sensitivity to detect an existing SNARC effect (Cipora & Wood, 2017). Finally, Bulut et al. (2023) conducted a lab study, while the present study was run online. However, it is unlikely that the difference in the results would originate from online or lab administration because the SNARC was successfully replicated in an online setup previously (e.g., Cipora et al., 2019a). Future investigations need to explore the reason for the difference between the results, but at present, lower power and lower measurement precision seem likely candidates for previous null effects. Also, one should consider that the Turkish SNARC was considerably smaller in size compared to the German one, therefore, the reason for previous null effect findings in the Turkish sample (Bulut et al., 2023) is likely a conjunction of lower power and the effect being smaller (consequently, even more difficult to detect with low power).

In Iranian culture, the regular group-level SNARC effect was significant in PJ and, even though, descriptively, being regular (i.e., negative slope), it was not significant in the MC task. Note that Bayesian statistics provided inconclusive evidence for both tasks in the Iranian group-level SNARC effect. Previous studies also demonstrated conflicting evidence for the SNARC effect at the group level in RL cultures; some reported a reverse SNARC effect (Shaki

et al., 2009, in Palestinian participants; Zebian, 2005, in Lebanese participants), others reported null effects (Shaki et al., 2009, in Israeli participants; Dehaene et al., 1993, Exp. 7, in Iranian participants; Lopiccolo & Chang, 2021, in Jordanian participants), and yet others reported a regular SNARC effect (Hochman et al., 2024; Zohar-Shai et al., 2017 in Israeli participants; Feldman et al., 2019 in Israeli children; Supplementary Material of Cipora et al., 2019a, in Iranian participants; Shaki & Gevers, 2011, in Israeli participants). Considering these and the present study's findings, it seems that more recent studies report regular SNARC effects more frequently, even in RL cultures, than earlier ones. How could a regular SNARC effect be explained in an RL culture? As we discussed above in cultural directionalities, a possible explanation is that LR directionality is innate and overwritten by culture (Adachi, 2014; Drucker & Brannon, 2014; McCrink & de Hevia, 2018; Rugani et al., 2015). Alternatively, instead of overwriting (by deleting the innate biases), cultural directionalities may diversify SNAs. This suggests that innate biases on SNAs are still present in adult life (see *The Brain's Asymmetric Frequency Tuning* hypothesis in Felisatti et al., 2020) and coexist with culture-dependent representations.

Even though culture clearly influences individuals' responses, this influence might not be strong enough to reverse the SNARC effect, but only weaken it (i.e., null or weak regular effects). Furthermore, global culture is dominated by LR cultures, as discussed above. With increasing globalization, populations from both LR and RL cultures are exposed to globalized cultural directional preferences. Support for this notion is that more recent studies (i.e., from a more globalized world) found a weakening of the effect, rather than a complete reversal. Moreover, some studies demonstrated the influence of experience with LR culture on the SNARC effect. For instance, Deheane et al. (1993) demonstrated a significant relationship between the time spent in France and the SNARC effect of Iranian participants; the more time in France, the stronger the regular SNARC effect. Furthermore, the level of SNARC effect was shown to be related to participants' status as either monoliterate (in RL languages) or biliterate (in both LR and RL languages) (see Zebian, 2005; Lopiccolo & Chang, 2021). In the present study, more than half of the Iranian participants reported speaking English as a foreign language and most stated they were somewhat familiar with LR cultures. Although these findings were in line with the globalization explanation of the absence of reverse SNARC effects in RL cultures, the familiarity of the Iranian participants with LR cultures and LR languages overshadows the influence of the cultural directionalities. Therefore, future studies should focus on SNAs among monolinguals/participants who have no/limited experience with LR cultures in general.

Finally, we should also note that Eastern Arabic numbers are written/read in LR fashion, and therefore, during a number-related task, LR activation of the number system might interfere with the responses. The above-mentioned factors may result in inconsistent directionalities in RL cultures, depending on the sample's characteristics, which in turn resulted in weakened, but not reversed, SNARC effect. A weak SNARC effect could be more susceptible to the power of the study, precision of measurement, and the task at hand, and consequently might produce inconsistent findings, which could explain the divergence of the findings of the SNARC effect in RL cultures. In sum, both the innateness of LR directionality and also, at a more individual level, experience with a predominantly LR global culture may explain why the SNARC effect is merely weakened, not reversed in an RL culture.

Markedness influencing response codes

Previously, effects of markedness on spatial association with response codes have frequently been reported. One such was the association of parity with response code (even-right and odd-left), which was first reported in German adults (Huber et al., 2015; Nuerk et al., 2004). Morphologically, German includes linguistic markedness in terms of even ("gerade") being the standard version of the word, and odd ("ungerade") being the opposite as marked by the prefix ("un"). This corresponds to formal markedness (see Nuerk et al., 2004) and is a strong determinant of the markedness grade. Therefore, we would expect a strong MARC effect in German culture. In the present study, there was some marginally significant ($p = 0.06$) standardized MARC effect, which may hint towards small markedness influence in the German sample. Importantly, Nuerk et al., (2004; see also Iversen et al., 2004, 2006) showed that the MARC effect is more consistent when the stimuli are presented in verbal notation, rather than in the Arabic notation used in this study.

There was no significant MARC effect in Turkish (see also Bulut et al., 2023) or Iranian samples. The morphological structure of markedness is lacking in Turkish ("tek"—odd, "çift"—even) and Farsi ("fard"—odd, "zoj"—even), which could be an explanation for the absence of MARC in these cultures. At the same time, the morphological structure of the language for parity has not always been found to affect the MARC effect (Cipora et al., 2019a, Supplementary Material).

Overall, the MARC effect with Arabic digits was not always replicated, as the effect is small in size and shows interindividual variability (see Huber et al., 2015). The current study adds to these inconsistent results.

Inter- and intraindividual prevalence of the SNARC effect across cultures

Intraindividual prevalence of the cognitive effects is typically overlooked. To investigate the reliability of an individual SNARC slope, Cipora et al., (2019b) suggested the H0 bootstrapping approach (see also Roth et al., 2024b). Using this method, we found that the reliable SNARC effect was observed most frequently in German (57% in PJ and MC), less in Turkish (38% in PJ and 46% in MC), and the least in Iranian (29% in PJ and 39% in MC). The proportion of reliable SNARC effect was found to be higher in German culture compared to the previously reported values from LR cultures (e.g., ≤ 45% in Cipora et al., 2019b; 42% in Hohol et al., 2022 for PJ, 42% in Hohol et al., 2020 Supplementary Material for MC). Although the reason for this higher proportion is not clear, it could be that participants performed both PJ and MC (see Supplementary Material Tables 15 and 16), whereas, in previous studies, participants only performed the PJ. Overall, these findings indicated that cultural directionalities, directional SNAs (i.e., the SNARC effect), and the prevalence of participants revealing a reliable SNARC effect systematically differed between cultures.

One important point we should address is that although the SNARC and CDPQ findings were comparable regarding the cultural differences, the CDPQ total scores did not correlate with individual SNARC slopes. This finding is not surprising considering that Roth et al. (2024b) demonstrated that a SNARC effect was present at the group level in their study, but was not stable over time for individual participants. Therefore, the absence of correlation between CDPQ and SNARC slopes is not necessarily a result of CDPQ's failure to predict the SNARC effect, but rather the poor stability of the individual SNARC slopes. A correlation between two measures in which one fluctuates unsystematically will be weak at best. In line with Roth et al. (2024b), we successfully captured the cross-cultural differences at the group level in both CDPQ and the SNARC, while correlations based on individual data did not show the same pattern. Note that further research is needed to fully establish reliability and validity of the CDPQ therefore conclusions should be made with caution. Nevertheless, the findings are promising and suggest that CDPQ could be a new tool to assess the cross-cultural differences in horizontal directional preferences.

Differences between PJ and MC

Examining the SNARC effect with PJ and MC allowed us to generalize the influence of cultural directionalities on both tasks. Comparing the slopes of these tasks and examining their relationship in a within-design is important in understanding whether they measure the same underlying construct. In the present study, there was no difference between the PJ and MC slopes in any of the cultures, and neither was any correlation observed. The correlation was similarly not significant between the PJ and MC slopes in the majority of the previous studies (see Cipora, 2014; Didino et al., 2019; Fattorini et al., 2015; Georges et al., 2017). Only two studies reported a significant correlation (Cheung et al., 2015; Cipora, 2014, only in an unimanual setup). It seems that even if there is a correlation between the two slopes, it is probably relatively weak, and requires a large sample to detect (e.g., Cheung et al., 2015 with 125 participants), which prompts a more fundamental question regarding the construct validity of SNA as an individual latent characteristic emerging across different tasks.

Finger counting habits cannot explain the sample differences in SNAs

Finger-counting habits are considered to influence the SNARC effect (Fischer, 2008; Fischer & Brugger, 2011). Although there is no conclusive evidence for this hypothesis (e.g., Prete & Tommasi, 2020; Cipora et al., 2019a, Supplementary Material; Hohol et al., 2022), we identified individuals' finger-counting direction to examine its influence on the SNARC effect, since such habits show variance across cultures (see Lindemann et al., 2011).

We were unable to replicate Fischer's (2008) findings in any of the cultures. In line with other studies (Fabbri, 2013; Hohol et al., 2022; Prete & Tommasi, 2020), the SNARC slope of left- and right-starters did not differ significantly. Bayesian analysis supported this lack of differences between the two groups. Although most participants reported stable finger-counting habits (88% in German, 77% in Turkish, and 95% in Iranian), no link was found between the stability of finger-counting habits and the level of the SNARC effect, similar to Hohol et al., (2020; 75% in Polish participants). It is important to note that in the present study, finger-counting habits were collected via self-report in an online setup, an approach that was shown to be inadequate in some studies (Lucidi & Thevenot, 2014; Morrissey & Hallett, 2018; Wasner et al., 2014). Therefore, more controlled lab studies in which participants physically count their fingers repeatedly are needed to accurately capture these habits and their stability for studies focusing on the relationship between finger-counting and the SNAs.

Limitations

We acknowledge some limitations of the study due to its online nature. Although the sample sizes were very similar in each groups at the beginning of the study, there was more data loss in Iran compared to other countries (e.g.,

some of the participants' data were not entirely stored, etc.). Although we believe that this issue should not be a problem with the current study's sample size (e.g., the minimum detectable effect size would be e.g., 0.41 in German-Iranian comparison for 0.80 power (alpha = 0.05, ratio = 130/75, two sample, two- sided, Welch test), findings should be interpreted with caution since the data loss was unequal across groups.

One should also note that the RTs may not be precise in online studies due to different browsers (Bridges et al., 2020) which might influence the measurement precision of the SNARC effect in the current study. At the same time, the SNARC effect has been successfully replicated in many online studies (e.g., Cipora et al., 2019a; Roth et al., 2024c, 2024d). We recorded the operating systems of the participants and no systematic difference was observed across samples (e.g., more than 75% of participants were using Windows 10 in all samples).

Conclusion

The present study supported our cultural directionality hypothesis, which stated that, in addition to reading direction, cultural directional preferences influence SNAs. Cultural directional preferences have previously been suggested to play a role in the emergence of SNAs in preliterate children, however, we provide evidence for the first time that such preferences influence SNAs in adults. The cultural environment's effect on cognition is highly complex and therefore should be examined from a comprehensive perspective. We suggest that examining cultural directionalities as a whole rather than focusing only on reading direction can help us understand the influence of culture on SNAs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00426-024-02038-4.

Acknowledgements We would like to thank Ceren Kaya, Andrey Andreev, Mohammad Reza Mohammad Hasani, and Mohammad Javad Rezazadeh for their contribution to data collection. We also would like to thank Dr. Simon Mumford for English proofreading the manuscript.

Author contribution All authors contributed to the manuscript directly and substantially.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was supported by the DFG project "Replicability of Fundamental Results on Spatial-Numerical Associations in Highly Powered Online Experiments (e-SNARC)" (NU 265/8–1) granted to Hans-Christoph Nuerk and Ulf-Dietrich Reips supporting Lilly Roth, with the assistance of Krzysztof Cipora as a cooperation partner. Krzysztof Cipora is supported by the UKRI Economic and Social Research Council (grant number ES/W002914/1). Merve Bulut was supported by TUBITAK 2214-A International Research Fellowship Program for PhD students.

Data availability The data, analyses, and preregistration are available in The Open Science Framework (OSF) at https://osf.io/dk4we/.

Declarations

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adachi, I. (2014). Spontaneous spatial mapping of learned sequence in chimpanzees: Evidence for a SNARC-like effect. *PLoS ONE, 9*(3), e90373. https://doi.org/10.1371/journal.pone.0090373

Bae, G. Y., Choi, J. M., Cho, Y. S., & Proctor, R. W. (2009). Transfer of magnitude and spatial mappings to the SNARC effect for parity judgments. *Journal of Experimental Psychology: Learning Memory, and Cognition, 35*(6), 1506–1521. https://doi.org/10.1037/a0017257

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ, 8*, e9414. https://doi.org/10.7717/peerj.9414

Bulf, H., de Hevia, M. D., & Macchi Cassia, V. (2016). Small on the left, large on the right: Numbers orient visual attention onto space in preverbal infants. *Developmental Science, 19*(3), 394–401.

Bulut, M., Hepdarcan, I., Palaz, E., Cetinkaya, H., & Dural, S. (2023). No SNARC effect among left- to-right readers: Evidence from a Turkish sample. *Advances in Cognitive Psychology, 19*(3), 224–236. https://doi.org/10.5709/acp-0394-x

Cheung C.-N., Ayzenberg V., Diamond R. F. L., Yousif S. R., Lourenco S. F. (2015). Probing the mental number line: A between-task analysis of spatial-numerical associations. In: Noelle D. C., Dale R., Warlaumont A. S., Yoshimi J., Matlock T., Jennings C. D. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 357–362). Lawrence Erlbaum.

Chokron, S., & De Agostini, M. (1995). Reading habits and line bisection: A developmental approach. *Cognitive Brain Research, 3*(1), 51–58. https://doi.org/10.1016/0926-6410(95)00018-6

Cipora, K. (2014). Between-task consistency of the SNARC effect. XXXIInd European Workshop on Cognitive Neuropsychology, Bressanone, Italy.

Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Goebel, S. M., Willmes, K., Pesenti, M., Schiltz, C., & Nuerk, H.-C. (2019b). *A minority pulls the sample mean: on the individual prevalence of robust group-level cognitive phenomena – the instance of the SNARC effect.* Open Science Framework [Preprint]. https://doi.org/10.31234/osf.io/bwyr3

Cipora, K., He, Y., & Nuerk, H.-C. (2020). The spatial–numerical association of response codes effect and math skills: Why related? *Annals of the New York Academy of Sciences, 1477*(1), 5–19. https://doi.org/10.1111/nyas.14355

Cipora, K., & Nuerk, H.-C. (2013). Is the SNARC effect related to the level of mathematics? No systematic relationship observed despite more power, more repetitions, and more direct assessment of arithmetic skill. *The Quarterly Journal of Experimental Psychology, 66*(10), 1974–1991. https://doi.org/10.1080/17470218.2013.772215

Cipora, K., Patro, K., & Nuerk, H.-C. (2018). Situated influences on spatial–numerical associations. In T. Hubbard (Ed.), *Spatial Biases in Perception and Cognition* (pp. 41–49). Cambridge University Press.

Cipora, K., Soltanlou, M., Reips, U.-D., & Nuerk, H.-C. (2019a). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behavior Research Methods, 51*(4), 1676–1692. https://doi.org/10.3758/s13428-019-01213-5

Cipora, K., & Wood, G. (2017). Finding the SNARC instead of hunting it: A 20*20 Monte Carlo investigation. *Frontiers in Psychology, 8*, 1194. https://doi.org/10.3389/fpsyg.2017.01194

de Hevia, M. D., Girelli, L., Addabbo, M., & Cassia, V. M. (2014). Human infants' preference for left-to-right oriented increasing numerical sequences. *PLoS ONE, 9*(5), e96412. https://doi.org/10.1371/journal.pone.0096412

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*(3), 371–396. https://doi.org/10.1037/0096-3445.122.3.371

Di Giorgio, E., Lunghi, M., Rugani, R., Regolin, L., Dalla Barba, B., Vallortigara, G., & Simion, F. (2019). A mental number line in human newborns. *Developmental Science, 22*, e12801. https://doi.org/10.1111/desc.12801

Didino, D., Breil, C., & Knops, A. (2019). The influence of semantic processing and response latency on the SNARC effect. *Acta Psychologica, 196*, 75–86. https://doi.org/10.1016/j.actpsy.2019.04.008

Didino, D., Brandtner, M., Glaser, M., & Knops, A.(2023). Probing the dual-route model of the SNARC effect by orthogonalizing processing speed and depth. *Experimental Psychology, 70*(1):1–13. https://doi.org/10.1177/00220221111435098

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice, 8*(1), 9–26. https://doi.org/10.1037/cns0000258

Drucker, C. B., & Brannon, E. M. (2014). Rhesus monkeys (Macaca mulatta) map number onto space. *Cognition, 132*(1), 57–67. https://doi.org/10.1016/j.cognition.2014.03.011

Ebersbach, M., Luwel, K., & Verschaffel, L. (2014). Further evidence for a spatial-numerical association in children before formal schooling. *Experimental Psychology, 61*(4), 323–329. https://doi.org/10.1027/1618-3169/a000250

Fabbri, M. (2013). Finger counting habits and spatial-numerical association in horizontal and vertical orientations. *Journal of Cognition and Culture, 13*(1–2), 95–110. https://doi.org/10.1163/15685373-12342086

Faghihi, N., Garcia, O., & Vaid, J. (2019). Spatial bias in figure placement in representational drawing: Associations with handedness and script directionality. *Laterality, 24*(5), 614–630. https://doi.org/10.1080/1357650X.2018.1561708

Faghihi, N., & Vaid, J. (2023). Reading/writing direction as a source of directional bias in spatial cognition: Possible mechanisms and scope. *Psychonomic Bulletin & Review, 30*(3), 843–862. https://doi.org/10.3758/s13423-022-02239-1

Fattorini, E., Pinto, M., Rotondaro, F., & Doricchi, F. (2015). Perceiving numbers does not cause automatic shifts of spatial attention. *Cortex, 73*, 298–316. https://doi.org/10.1016/j.cortex.2015.09.007

Feldman, A., Oscar-Strom, Y., Tzelgov, J., & Berger, A. (2019). Spatial–numerical association of response code effect as a window to mental representation of magnitude in long-term memory among Hebrew-speaking children. *Journal of Experimental Child Psychology, 181*, 102–109. https://doi.org/10.1016/j.jecp.2019.01.001

Felisatti, A., Laubrock, J., Shaki, S., & Fischer, M. H. (2020). A biological foundation for spatial–numerical associations: The brain's asymmetric frequency tuning. *Annals of the New York Academy of Sciences, 1477*, 44–53. https://doi.org/10.1111/nyas.14418

Fias, W., Brysbaert, M., Geypens, F., & D'Ydewalle, G. (1996). The importance of magnitude information in numerical processing: Evidence from the SNARC effect. *Mathematical Cognition, 2*(1), 95–110. https://doi.org/10.1080/135467996387552

Fischer, M. H. (2008). Finger counting habits modulate spatial-numerical associations. *Cortex, 44*(4), 386–392. https://doi.org/10.1016/j.cortex.2007.08.004

Fischer, M. H., & Brugger, P. (2011). When digits help digits: Spatial, numerical associations point to finger counting as prime example of embodied cognition. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2011.00260

Fischer, M. H., Shaki, S., & Cruise, A. (2009). It takes just one word to quash a SNARC. *Experimental Psychology, 56*(5), 361–366. https://doi.org/10.1027/1618-3169.56.5.361

Fitousi, D., Shaki, S., & Algom, D. (2009). The role of parity, physical size, and magnitude in numerical cognition: the SNARC effect revisited. *Attention, perception & psychophysics, 71*(1), 143–155. https://doi.org/10.3758/APP.71.1.143

Garaizar, P., & Reips, U.-D. (2019). Best practices: Two web browser-based methods for stimulus presentation in behavioral experiments with high resolution timing requirements. *Behavior Research Methods, 51*, 1441–1453. https://doi.org/10.3758/s13428-018-1126-4

Georges, C., Hoffmann, D., & Schiltz, C. (2017). Mathematical abilities in elementary school: do they relate to number–space associations? *Journal of Experimental Child Psychology, 161*, 126–147. https://doi.org/10.1016/j.jecp.2017.04.011

Gevers, W., Santens, S., Dhooge, E., Chen, Q., van den Bossche, L., Fias, W., & Verguts, T. (2010). Verbal-spatial and visuospatial coding of number-space interactions. *Journal of Experimental Psychology: General, 139*(1), 180–190. https://doi.org/10.1037/a0017688

Gevers, W., Verguts, T., Reynvoet, B., Caessens, B., & Fias, W. (2006). Numbers and space: a computational model of the SNARC effect. *Journal of Experimental Psychology Human Perception and Performance, 32*(1):32–44. https://doi.org/10.1037/0096-1523.32.1.32

Göbel, S. M., McCrink, K., Fischer, M. H., & Shaki, S. (2018). Observation of directional storybook reading infuences young children's counting direction. *Journal of Experimental Child Psychology, 166*, 49–66. https://doi.org/10.1016/j.jecp.2017.08.001

Hung, Y. H., Hung, D. L., Tzeng, O. J. L., & Wu, D. H. (2008). Flexible spatial mapping of different notations of numbers in Chinese readers. Cognition, 106 (3), 1441-1450. https://doi.org/10.1016/j.cognition.2007.04.017

Hochman, S., Havedanloo, R., Heysieattalab, S., & Soltanlou, M. (2024, September 5). How does language modulate the association between number and space? A registered report of a cross-cultural study of the SNARC effect. https://doi.org/10.31234/osf.io/sme3z

Hoffmann, D., Hornung, C., Martin, R., & Schiltz, C. (2013). Developing number-space associations: SNARC effects using a color discrimination task in 5-year-olds. *Journal of Experimental Child*

Psychology, 116(4), 775–791. https://doi.org/10.1016/j.jecp.2013.07.013

Hohol, M., Willmes, K., Nęcka, E., Brozek, B., Nuerk, H.-C., & Cipora, K. (2020). Professional mathematicians do not differ from others in the symbolic numerical distance and size effects. *Scientific Reports*. https://doi.org/10.1038/s41598-020-68202-z

Hohol, M., Wołoszyn, K., & Cipora, K. (2022). No fingers, no SNARC? Neither the finger counting starting hand, nor its stability robustly affect the SNARC effect. *Acta Psychologica, 230*, 103765. https://doi.org/10.1016/j.actpsy.2022.103765

Hohol, M., Wołoszyn, K., Nuerk, H.-C., & Cipora, K. (2018). A large-scale survey on finger counting routines, their temporal stability and flexibility in educated adults. *PeerJ, 6*, e5878. https://doi.org/10.7717/peerj.5878

Huber, S., Klein, E., Graf, M., Nuerk, H.-C., Moeller, K., & Willmes, K. (2015). Embodied markedness of parity? Examining handedness effects on parity judgments. *Psychological Research Psychologische Forschung, 79*, 963–977. https://doi.org/10.1007/s00426-014-0626-9

Ito, Y., & Hatta, T. (2004). Spatial structure of quantitative representation of numbers: Evidence from theSNARC effect. *Memory & Cognition, 32*(4), 662–673. https://doi.org/10.3758/BF03195857

Iversen, W., Nuerk, H.-C., Jager, L., & Willmes, K. (2006). The influence of an external symbol system on number parity representation, or what's odd about 6? *Psychonomic Bulletin & Review, 13*, 730–736. https://doi.org/10.3758/BF03193988

Iversen, W., Nuerk, H.-C., & Willmes, K. (2004). Do signers think differently? The processing of number parity in deaf participants. *Cortex, 40*, 176–178. https://doi.org/10.1016/s0010-9452(08)70940-7

Kebbe, H., & Vinter, A. (2013). How culture, age, and manual dominance affect directionality in drawing side view objects. *Journal of Cross Cultural Psychology, 44*, 160–172. https://doi.org/10.1177/0022022111435098

Lieblich, A., Ninio, A., & Kugelmass, S. (1975). Developmental trends in directionality of drawing in Jewish and Arab Israeli children. *Journal of Cross Cultural Psychology, 6*, 504–511. https://doi.org/10.1177/002202217564013

Lindemann, O., Alipour, A., & Fischer, M. H. (2011). Finger counting habits in middle eastern and Western individuals: An online survey. *Journal of Cross-Cultural Psychology, 42*(4), 566–578. https://doi.org/10.1177/0022022111406254

Lopiccolo, D., & Chang, C. B. (2021). Cultural factors weaken but do not reverse left-to-right spatial biases in numerosity processing: Data from Arabic and English monoliterates and Arabic-English biliterates. *PLoS ONE, 16*(12), e0261146. https://doi.org/10.1371/journal.pone.0261146

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology, 16*(1), 149–157.

Lucidi, A., & Thevenot, C. (2014). Do not count on me to imagine how I act: Behavior contradicts questionnaire responses in the assessment of finger counting habits. *Behavior Research Methods, 46*(4), 1079–1087. https://doi.org/10.3758/s13428-014-0447-1

Maass, A., & Russo, A. (2003). Directional bias in the mental representation of spatial events: Nature or culture? *Psychological Science, 14*(4), 296–301. https://doi.org/10.1111/1467-9280.14421

McCrink, K., & de Hevia, M. D. (2018). From innate spatial biases to enculturated spatial cognition: The case of spatial associations in number and other sequences. *Frontiers in Psychology, 9*, 415. https://doi.org/10.3389/fpsyg.2018.00415

McCrink, K., & Opfer, J. E. (2014). Development of spatial–numerical associations. *Current Directions in Psychological Science, 23*(6), 439–445. https://doi.org/10.1177/0963721414549751

Morrissey, K., & Hallett, D. (2018). Cardinal and ordinal aspects of finger-counting habits predict different individual differences in embodied numerosity. *Journal of Numerical Cognition, 4*(3), 613–635. https://doi.org/10.5964/jnc.v4i3.138

Nuerk, H.-C., Iversen, W., & Willmes, K. (2004). Notational modulation of the SNARC and the MARC (Linguistic Markedness of Response Codes) effect. *The Quarterly Journal of Experimental Psychology Section A, 57*(5), 835–863. https://doi.org/10.1080/02724980343000512

Nuerk, H.-C., Patro, K., Cress, U., Schild, U., Friedrich, C. K., & Gobel, S. M. (2015). How space-number associations may be created in preliterate children: six distinct mechanisms. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.00215

Nuerk, H.-C., Wood, G., & Willmes, K. (2005). The universal SNARC effect. *Experimental Psychology, 52*(3), 187–194. https://doi.org/10.1027/1618-3169.52.3.187

Patro, K., Nuerk, H.-C., & Cress, U. (2016a). Mental number line in the preliterate brain: The role of early directional experiences. *Child Development Perspectives, 10*(3), 172–177. https://doi.org/10.1111/cdep.12179

Patro, K., Fischer, U., Nuerk, H.-C., & Cress, U.(2016b). How to rapidly construct a spatial–numerical representation in preliterate children (at least temporarily). *Developmental Science, 19* (1), 126–144. https://doi.org/10.1111/desc.12296

Patro, K., & Haman, M. (2012). The spatial-numerical congruity effect in preschoolers. *Journal of Experimental Child Psychology, 111*, 534–542. https://doi.org/10.1016/j.jecp.2011.09.006

Patro, K., Nuerk, H.-C., & Cress, U. (2015). Does your body count? Embodied influences on the preferred counting direction of preschoolers. *Journal of Cognitive Psychology, 27*(4), 413–425. https://doi.org/10.1080/20445911.2015.1008005

Pitt, B., & Casasanto, D. (2020). The correlations in experience principle: How culture shapes concepts of time and number. *Journal of Experimental Psychology: General, 149*(6), 1048–1070. https://doi.org/10.1037/xge0000696

Prete, G., & Tommasi, L. (2020). Exploring the interactions among SNARC effect, finger counting direction and embodied cognition. *PeerJ*. https://doi.org/10.7717/peerj.9155

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin, 132*(3), 416–442. https://doi.org/10.1037/0033-2909.132.3.416

Rashidi-Ranjbar, N., Goudarzvand, M., Jahangiri, S., Brugger, P., & Loetscher, T. (2014). No horizontal numerical mapping in a culture with mixed reading habits. *Frontiers in Human Neuroscience, 8*, 72. https://doi.org/10.3389/fnhum.2014.00072

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49*, 243–256. https://doi.org/10.1026/1618-3169.49.4.243

Reips, U.-D. (2021). Web-based research in psychology: A review. *Zeitschrift Für Psychologie, 229*(4), 198–213.

Rinaldi, L., Di Luca, S., Henik, A., & Girelli, L. (2014). Reading direction shifts visuospatial attention: An Interactive Account of attentional biases. *Acta Psycholiga, 151*, 98–105. https://doi.org/10.1016/j.actpsy.2014.05.018

Roth, L., Caffier, J., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2024c, in press). True colors SNARCing: Semantic number processing is highly automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Preprint at: https://osf.io/preprints/psyarxiv/aeyn8

Roth, L., Caffier, J., Reips, U.-D., Nuerk, H.-C., Overlander, A. T., & Cipora, K. (2024d, in press). One and only SNARC? Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude. *Royal Society Open Science*. Preprint at: https://osf.io/79zsy/

Roth, L., Cipora, K., Overlander, A. T., Nuerk, H.-C., & Ulf-Dietrich, R. (2024a). *Shape of SNARC: How task-dependent are Spatial-Numerical Associations? A highly powered online experiment* [preprint of Stage-1 Registered Report, submitted]. https://osf.io/4wpv6/

Roth, L., Jordan, V., Schwarz, S., Willmes, K., Nuerk, H.-C., van Dijck, J.P., & Cipora, K. (2024b). Don't SNARC me now! Intraindividual variability of cognitive phenomena – insights from the Ironman paradigm. *Cognition 248,* 105781. https://doi.org/10.31234/osf.io/g4bjq

Rugani, R., Vallortigara, G., Priftis, K., Regolin, L. (2015). Animal cognition. Number-space mapping in the newborn chick resembles humans' mental number line. *Science 30*;347(6221), 534–536. https://doi.org/10.1126/science.aaa1379

Schild, U., Steil, J. N., Ulrich, R., & Friedrich, C. (2022). *Children anticipate past events to the left and future events to the right: Evidence from eye movements and time-space compatibility effects.* https://doi.org/10.31234/osf.io/cjzd4

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017). Switching between multiple codes of SNARC-like associations: Two conceptual replication attempts with anodal tDCS in sham-controlled cross-over design. *Frontiers in Neuroscience, 11*, 654. https://doi.org/10.3389/fnins.2017.00654

Shaki, S., & Fischer, M. H. (2024). How do numbers shift spatial attention? Both processing depth and counting habits matter. *Journal of Experimental Psychology. General, 153*(1), 171–183. https://doi.org/10.1037/xge0001493

Shaki, S., Fischer, M., & Göbel, S. (2012). Direction counts: A comparative study of spatially directional counting biases in cultures with different reading directions. *Journal of Experimental Child Psychology, 112*(2), 275–281. https://doi.org/10.1016/j.jecp.2011.12.005

Shaki, S., & Fischer, M. H. (2008). Reading space into numbers – a cross-linguistic comparison of the SNARC effect. *Cognition, 108*(2), 590–599

Shaki, S., Fischer, M., & Petrusic, W. (2009). Reading habits for both words and numbers contribute to the SNARC effect. *Psychonomic Bulletin & Review, 16*(2), 328–331. https://doi.org/10.3758/PBR.16.2.328

Shaki, S., & Gevers, W. (2011). Cultural characteristics dissociate magnitude and ordinal information processing. *Journal of Cross-Cultural Psychology, 42*(4), 639–650. https://doi.org/10.1177/0022022111406100

Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology, 23*, 515–557. https://doi.org/10.1016/0010-0285(91)90005-9

Vaid, J. (1995). Script directionality affects nonlinguistic performance: Evidence from Hindi and Urdu. In I. Taylor & D. R. Olson (Eds.), *Scripts and Literacy* (pp. 295–310). Springer.

van Dijck, J. P., & Fias, W. (2011). A working memory account for spatial–numerical associations. *Cognition, 119*, 114–119. https://doi.org/10.1016/j.cognition.2010.12.013

van Dijck, J. P., Fias, W., & Cipora, K. (2022). Spatialization in working memory and its relation to math anxiety. *Annals of the New York Academy of Sciences, 1512*(1), 192–202. https://doi.org/10.1111/nyas.14765

Wasner, M., Moeller, K., Fischer, M. H., & Nuerk, H.-C. (2014). Aspects of situated cognition in embodied numerosity: The case of finger counting. *Cognitive Processing, 15*(3), 317–328. https://doi.org/10.1007/s10339-014-0599-z

Winter, B., Matlock, T., Shaki, S., & Fischer, M. H. (2015). Mental number space in three dimensions. *Neuroscience and Biobehavioral Reviews, 57*, 209–219. https://doi.org/10.1016/j.neubiorev.2015.09.005

Wood, G., Willmes, K., Nuerk, H.-C., & Fischer, M. H. (2008). On the cognitive link between space and number: A meta-analysis of the SNARC effect. *Psychology Science Quarterly, 50*(4), 489–525.

Yılmaz, H. (2011). Learning to read (again): The social experiences of Turkey's 1928 Alphabet Reform. *International Journal of Middle East Studies, 43*(4), 677–697. https://doi.org/10.1017/S0020743811000900

Zebian, S. (2005). Linkages between number concepts, spatial thinking, and directionality of writing: The SNARC effect and the reverse SNARC effect in English and Arabic monoliterates, biliterates, and illiterate Arabic speakers. *Journal of Cognition and Culture, 5*(1–2), 165–190. https://doi.org/10.1163/1568537054068660

Zohar-Shai, B., Tzelgov, J., Karni, A., & Rubinsten, O. (2017). It does exist! A left-to-right spatial–numerical association of response codes (SNARC) effect among native Hebrew speakers. *Journal of Experimental Psychology: Human Perception and Performance, 43*(4), 719–728. https://doi.org/10.1037/xhp0000336

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A3.2. Supplementary materials for Study 3

These supplementary materials belong to the following publication (see Appendix A3.1.):

Bulut, M., Roth, L., Bahreini, N., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2025). One direction? Cultural aspects of the mental number line beyond reading direction. *Psychological Research, 89*, Article 37. https://doi.org/10.1007/s00426-024-02038-4

Supplementary Materials

Table 1

Detailed information about the languages spoken by bilingual individuals. The datasets from bilingual participants whose second language is written in another direction from the first one (left-to-right for Iran and right-to-left for Germany and Turkey) or whose second language belongs to the Middle Eastern (for Germany) or Western culture (for Turkey and Iran) were excluded

The Sample	Second Native Language	Opposite direction	Middle Eastern (for German) or Western (for Turkish and Iranian) culture	Exclusion	*N*
German	Turkish	no	Middle Eastern	**excluded**	**3**
German	Arabic	yes	Middle Eastern	**excluded**	**1**
German	Spanish	no	-	not excluded	2
German	Greek	no	-	not excluded	2
German	Vietnamese	no	-	not excluded	1
German	Croatian	no	-	not excluded	1
German	Polish	no	-	not excluded	2
German	Bulgarian	no	-	not excluded	1
German	Portuguese	no	-	not excluded	1
German	Russian	no	-	not excluded	1
German	Italian	no	-	not excluded	1
Turkish	German	no	Western	**excluded**	**1**
Turkish	Dutch	no	Western	**excluded**	**1**

Turkish	English	no	Western	**excluded**	**1**
Iranian	Turkish	yes	-	**excluded**	**4**
Iranian	English	yes	Western	**excluded**	**3**
Iranian	Kurdish	no	-	not excluded	1
Iranian	Arabic	no	-	not excluded	1
Iranian	Sangsari	no	-	not excluded	1

Table 2

Foreign languages spoken by the participants. None of the datasets of participants were excluded based on foreign language.

The Language	The Direction	The Sample	N
English	LR	German	126
		Turkish	96
		Iranian	48
Spanish	LR	German	41
		Turkish	1
		Iranian	1
German	LR	Turkish	11
		Iranian	4
French	LR	German	52
		Turkish	1
Arabic	RL	German	3
		Turkish	1
		Iranian	23
Italian	LR	German	12
		Turkish	1
Turkish	LR	German	1
		Iranian	2

Russian	LR	German Turkish Iranian	1 2 1
Japanese	LR[a]	German Turkish	2 3
Dutch	LR	German	3
Farsi	RL	German	1
Mandarin Chinese	LR	German	1
Korean	LR	German	2
Romanian	LR	German	1
Greek	LR	German	1
Norwegian	LR	German	2
Hindu/Urdu	LR/RL	German	1

[a]Japanese is also written from top-to-bottom

The impact of foreign language use

Only three participants in the German and one in the Turkish sample reported knowing a foreign language with RL direction. Forty-six Iranian participants reported knowing a foreign language with LR direction. Therefore, there was only considerable data in the Iranian sample to check whether foreign language use that is in the opposite direction of the native language had any impact on the SNARC slopes. Forty-eight participants reported that they speak English as a foreign language. The majority of them (i.e., 40 participants) reported that they were either actively using English (20 participants) or they had used it in the last five years (20 participants). Eight participants reported that they had not used English actively in the last five years. We considered the latter ones as non-users and excluded their data from this analysis. For the

remaining participants, the correlation analyses showed no significant relationship between foreign language use and the SNARC slopes in any of the tasks (PJ: handwriting: $r = .27$, $p = .089$, typing: $r = -.07$, $p = .896$, spoken: $r = .02$, $p = .883$; MC: handwriting: $r = .15$, $p = .370$, typing: $r = .07$, $p = .664$, spoken: $r = .09$, $p = .599$).

Table 3

Familiarity of participants with the Western (for the Turkish and Iranian samples) and Middle Eastern cultures (for the German sample)

Sample	Not familiar (*N*)	A bit familiar (*N*)	Very familiar (*N*)
German	53.85% (70)	44.62% (58)	0.77% (1)
Turkish	7.14% (8)	60.71% (68)	32.14% (36)
Iranian	2.67% (2)	78.67% (59)	12.50% (14)

Note. 1 participant in the German sample did not answer to this question

Task- and block-order influence on the SNARC effect

Independent samples *t*-tests showed that in none of the tasks and none of the samples, the task order had an effect on the unstandardized or standardized SNARC slopes (see Table 4).

Another series of independent sample *t*-tests indicated that there was no significant effect of block order on PJ unstandardized SNARC slopes (see Table 5). Interestingly, in the German sample, there was a significant effect of block order ($t(113.97) = 3.05$, $p = .003$, $BF_{10} = 12.39$, *adj*. $p = .017$) on MC SNARC slopes. Based on this analysis, participants who performed the incompatible blocks first showed a stronger SNARC effect ($M = -49.37$) in the MC task compared to participants who performed the compatible blocks first ($M = -16.30$). Similar findings were obtained for standardized SNARC slopes in the MC task (German: $t(128) = 2.87$, p

= 0.005, BF_{10} = 7.30, *adj. p* = .029). This difference was not observed in other samples (see Table 5)

Table 4

Two-sided independent-samples t-test findings indicating that there was no task-order influence on the unstandardized and standardized SNARC slopes

Measure	Sample	Task	*t*	*df*	*p*	BF_{10}
SNARC	German	PJ	0.81	125.41	.421	0.25*
		MC	-0.99	102.23	.323	0.30*
	Turkish	PJ	0.69	109.94	.489	0.25*
		MC	-1.14	103.54	.258	0.36
	Iranian	PJ	-1.44	55.23	.156	0.59
		MC	0.43	71.43	.666	0.26*
Standardized SNARC	German	PJ	0.58	125.15	.560	0.22*
		MC	0.24	107.85	.813	0.19*
	Turkish	PJ	-0.44	109.55	.658	0.22*
		MC	-1.83	109.86	.071	0.88
	Iranian	PJ	-1.24	51.71	.220	0.49
		MC	-0.23	45.89	.818	0.25*

***** $p < .05$ indicating a significant finding (frequentist), $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis, and $BF_{10} > 3$ indicating a conclusive finding for the alternative hypothesis (Bayesian)

Table 5

Two-sided independent-samples t-test findings indicating the block-order influence on the unstandardized and standardized SNARC slopes

Measure	Sample	Task	*t*	*df*	*p*	BF_{10}
SNARC	German	PJ	-0.16	127.93	.872	0.19*
		MC	3.05	113.97	.017[a]*	12.39*
	Turkish	PJ	0.28	108.68	.781	0.21*
		MC	0.96	105.86	.341	0.29*
	Iranian	PJ	0.62	64.38	.537	0.28*
		MC	1.00	48.98	.320	0.38
Standardized SNARC	German	PJ	0.53	126.70	.598	0.21*
		MC	2.87	128	.029[a]*	7.30*
	Turkish	PJ	0.19	105.90	.852	0.21*
		MC	1.25	109.86	.215	.038
	Iranian	PJ	0.41	72.53	.684	0.26*
		MC	2.46	72.38	.08[a]	3.09*

[a] p is corrected with Holm-Bonferroni for 6 tests

* $p < .05$ indicating a significant finding (frequentist), $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis, and $BF_{10} > 3$ indicating a conclusive finding for the alternative hypothesis (Bayesian)

Table 6

Finger counting habits across samples

Sample (*N*)	Left-starters (*N*)	Right-starters (*N*)	No preference (*N*)
German (130)	43.85% (57)	50.00% (65)	6.15% (8)
Turkish (112)	55.36% (62)	35.71% (40)	8.93% (10)
Iranian (75)	26.03% (19)	64.38% (47)	9.59% (7)

Note. Two participants in the Iranian sample did not respond to the finger counting questions.

Table 7

Finger counting habit and its stability among left- and right- starters across samples

Sample	Left-starters			Right-starters		
	Always (*N*)	Usually (*N*)	Not stable (*N*)	Always (*N*)	Usually (*N*)	Not stable (*N*)
German	43.86% (25)	43.86% (25)	12.28% (7)	43.08% (28)	43.08% (28)	13.85% (9)
Turkish	40.32% (25)	37.10% (23)	22.58% (14)	37.50% (15)	50.00% (20)	12.50% (5)
Iranian	26.32% (5)	68.42% (13)	5.26% (1)	40.43% (19)	55.32% (26)	4.25% (2)

Table 8

Two-sided Jonckheere-Terpstra tests findings comparing the unstandardized SNARC slopes based on the finger-counting stability among left- and right-starters

Sample	Task	Starting Hand	T_{JT}	p	BF_{10}
German	PJ	Left	532	.537	0.43
		Right	654	.881	0.36
	MC	Left	498	.854	0.41
		Right	697	.526	0.36
Turkish	PJ	Left	570	.528	0.32*
		Right	276	.327	0.63
	MC	Left	444	.252[a]	2.44
		Right	290	.159	0.75
Iranian	PJ	Left	25	.142	-
		Right	262	.541	0.60
	MC	Left	38	.833	-
		Right	294	.964	0.98

Note 1. BF's include only always and no preference group.
Note 2. There was not enough observation for BF in Iranian left starters.
[a] p is corrected with Holm-Bonferroni correction for 12 tests
***** $p < .05$ indicating a significant finding (frequentist), $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis, and $BF_{10} > 3$ indicating a conclusive finding for the alternative hypothesis (Bayesian)

Table 9

The SNARC effect and the MARC effect in PJ task without applying bilingualism and location filter

Measure	Sample	Mean Slope (SD)	*t*	*p*	BF_{10}
SNARC	German	-5.91 (6.67)	-10.22	< .001*	2.60e+15*
	Turkish	-3.50 (8.54)	-4.40	< .001*	639.77*
	Iranian	-1.96 (7.04)	-2.49	.015*	2.19
Standardized SNARC	German	-0.42 (0.45)	-10.65	< .001*	2.91e+16*
	Turkish	-0.21 (0.45)	-4.98	< .001*	5987.48*
	Iranian	-0.12 (0.47)	-2.23	.029*	1.26
MARC	German	-7.53 (68.04)	-1.28	.204	0.22*
	Turkish	0.56 (85.92)	0.07	.944	0.10*
	Iranian	-8.34 (80.21)	-0.93	.355	0.19*
Standardized MARC	German	-0.12 (0.84)	-1.72	.088	0.40
	Turkish	0.01 (0.85)	0.07	.943	0.10*
	Iranian	-0.11 (0.82)	-1.24	.217	0.26*

Note 1. All tests were two-sided and against zero
Note 2. *df* was 132 for the German, 115 for the Turkish, and 79 for the Iranian sample
***** $p < .05$ indicating a significant finding (frequentist), $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis, and $BF_{10} > 3$ indicating a conclusive finding for the alternative hypothesis (Bayesian)

Table 10

The SNARC effect in MC task without applying bilingualism and location filter

Measure	Sample	Mean Slope (SD)	*t*	*p*	BF_{10}
SNARC	German	-32.45 (63.30)	-5.91	< .001*	393308.50*
	Turkish	-26.52 (74.98)	-3.79	< .001*	77.72*
	Iranian	-20.90 (102.21)	-1.83	.071	0.60
Standardized SNARC	German	-0.34 (0.61)	-6.34	< .001*	3.50e+16*
	Turkish	-0.22 (0.60)	-3.96	< .001*	2773.61*
	Iranian	-0.14 (0.58)	-2.17	.033	1.26

Note 1. All tests were against zero
Note 2. df was 132 for the German, 115 for the Turkish, and 79 for the Iranian sample
***** $p < .05$ indicating a significant finding (frequentist), $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis, and $BF_{10} > 3$ indicating a conclusive finding for the alternative hypothesis (Bayesian)

Table 11

The comparison of the SNARC effect across samples without applying bilingualism and location filter

Task	JT[a]	*p*
PJ	21027	< .001*
MC	18974	.064

Note. One-sided JT was performed for the alternative hypothesis of German < Turkish < Iranian order for unstandardized SNARC slopes
[a] $BF_{10} = 2.59$ in German-Turkish and $BF_{10} = 0.36$ in Turkish-Iranian comparison in PJ task
[a] $BF_{10} = 0.17$* in German-Turkish and $BF_{10} = 0.17$* in Turkish-Iranian comparison in MC task
***** $p < .05$ indicating a significant finding (frequentist), $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis, and $BF_{10} > 3$ indicating a conclusive finding for the alternative hypothesis (Bayesian)

Table 12

The comparison of the SNARC effect between left- and right-starters with two-sided independent samples t-test without applying bilingualism and location filter

Task	Sample	*t*	*df*	*p*	BF_{10}
PJ	German	0.66	122.97	.508	0.23*
	Turkish	-0.52	68.75	.605	0.24*
	Iranian	0.03	34.47	.974	0.27*
MC	German	-1.10	117.99	.275	0.33*
	Turkish	0.18	86.37	.858	0.22*
	Iranian	0.74	34.06	.466	0.33*

***** $p < .05$ indicating a significant finding (frequentist), $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis, and $BF_{10} > 3$ indicating a conclusive finding for the alternative hypothesis (Bayesian)

Table 13

The comparison of the SNARC effect based on the stability of finger counting habits with two-sided Jonkheere Terpsta test without applying bilingualism and location filter

Sample	Task	Starting Hand	T_{JT}	p
German	PJ	Left	532	.514
		Right	713	.877
	MC	Left	498	.855
		Right	754	.542
Turkish	PJ	Left	592	.377
		Right	276	.349
	MC	Left	458	.144[a]
		Right	290	.170
Iranian	PJ	Left	25	.176
		Right	312	.428
	MC	Left	38	.798
		Right	346	.843

[a] p is corrected with Holm-Bonferroni

***** $p < .05$ indicating a significant finding

Table 14

The comparison of SNARC slopes of PJ and MC tasks (calculated with continuous predictor) with two-sided paired-samples t-test without applying bilingualism and location filter

Sample	*t*	*df*	*p*	BF_{10}
German	0.01	132	.992	0.10*
Turkish	-0.87	114	.385	0.15*
Iranian	-1.02	79	.311	0.20*

* $p < .05$ indicating a significant finding (frequentist), $BF_{10} < 1/3$ indicating a conclusive finding for the null hypothesis, and $BF_{10} > 3$ indicating a conclusive finding for the alternative hypothesis (Bayesian)

Table 15

Individual prevalence of the reliable SNARC effect calculated with H0 bootstrapping in PJ based on task order

Task Order	Sample (N)	Reliable SNARC (*N*)	Reliable reverse SNARC (*N*)	Unreliable SNARC (*N*)
MC-PJ	German (58)	25.38% (33)	2.31% (3)	16.92% (22)
	Turkish (59)	18.75% (21)	5.36% (6)	28.57% (32)
	Iranian (27)	12.00% (9)	2.67% (2)	21.33% (16)
PJ-MC	German (72)	31.54% (41)	2.30% (3)	21.53 % (28)
	Turkish (53)	18.75% (21)	6.25% (7)	22.32% (25)
	Iranian (48)	17.33% (13)	9.33% (7)	37.33% (28)

Table 16

Individual prevalence of the reliable SNARC effect calculated with H0 bootstrapping in MC based on task order

Task Order	Sample (N)	Reliable SNARC (*N*)	Reliable reverse SNARC (*N*)	Unreliable SNARC (*N*)
MC-PJ	German (58)	24.62% (32)	7.69% (10)	12.31% (16)
	Turkish (59)	25.89% (29)	6.25% (7)	20.54% (23)
	Iranian (27)	14.67% (11)	6.67% (5)	14.67% (11)
PJ-MC	German (72)	32.31% (42)	8.46% (11)	14.62 % (19)
	Turkish (53)	20.54% (23)	13.39% (15)	13.39% (15)
	Iranian (48)	24.00% (18)	14.67% (11)	25.33% (19)

The block-order influence on the MARC effect in PJ

The influence of the block order on the MARC effect was examined by performing independent samples *t*-tests on the unstandardized MARC slopes across all samples. Results revealed that performing incongruent blocks first (M = -25.25 in the German; M = -19.26 in the Turkish, and M = -24.62 in the Iranian samples) revealed stronger MARC effects compared to performing congruent blocks first (M = 7.96 in the German; M = 20.32 in the Turkish, and M = 16.43 in the Iranian samples) in PJ (German: $t(127.52) = 2.85$, $p = .005$, *adj.* $p = .015$; Turkish: $t(93.89) = 2.50$, $p = .014$, *adj.* $p = .028$; Iranian: $t(63.61) = 2.50$, $p = .015$, *adj.* $p = .028$). Interestingly, the MARC effect observed among participants who performed the incongruent blocks first was significant at the group level in the German sample as indicated by a one-sample *t*-test against zero ($t(63) = -3.19$, $p = .002$, *adj. p.* = .013). Other MARC effects were not

significant (German sample: congruent blocks first: $t(65) = 0.24$, $p = .814$; Turkish sample congruent blocks first: $t(63) = 2.09$, $p = .041$, *adj. p.* = .162, incongruent blocks first: $t(46) = -1.54$, $p = .130$; Iranian sample congruent blocks first: $t(35) = 1.22$, $p = .229$, incongruent blocks first, $t(38) = -2.62$, $p = .013$, *adj. p.* = .063).

Reliability of the SNARC effect

We calculated the split-half reliability of the SNARC effect in each sample by using the odd-even method based on the presentation order of trials (Cipora et al., 2019a). Unstandardized slopes were re-calculated for odd and even series separately for each participant. The correlation analysis was performed on these recalculated scores. The Pearson correlation coefficients were adjusted for test length by using Spearman-Brown correction. For the PJ task, the reliability of the SNARC effect was .49 in German, .61 in Turkish, and .59 in the Iranian samples. These reliabilities were comparable to the ones calculated in online PJ tasks (.40 and .43 in Hohol et al., 2022; Cipora et al., 2019a, respectively) but rather low compared to in-lab SNARC studies with PJ tasks (.82, .70, .75, and .55 in Cipora et al. 2016; Cipora & Nuerk, 2013; Fattarini et al., 2015; Georges et al., 2017, respectively). For the MC tasks, the reliability of the SNARC effect was .90 in the German, .91 in the Turkish, and .93 in the Iranian sample. To the best of our knowledge, there is no previous reliability report of an MC task. Based on the typical psychometric criteria, we can conclude that the reliability of the MC task was quite high across all cultures. Note that the high reliability in MC could also be a result of the categorical predictor being used in SNARC slope calculations (for a similar calculation for MARC effect reliability see Cipora et al., 2019a which also reports reliability for MARC as .91).

The association of good and bad with the horizontal space

Additionally, Fisher's exact test showed that there was a significant association between the sample and the horizontal space of good and bad, $p < .001$. Bonferroni-corrected post-hoc comparisons revealed that the good-right bad-left association was more frequently selected in the Iranian compared to the German ($p < .001$) and Turkish ($p < .001$) samples. Also, the good-right bad-left association was more frequently selected in the Turkish compared to the German sample ($p = .030$).

Table 17

Operating systems used by participants across samples

Sample (*N*)	Windows 10 (*N*)	Mac OS X 10.15.17 (*N*)	Other (*N*)
German (130)	76.92% (100)	18.46% (24)	4.62% (6)
Turkish (112)	82.14% (92)	13.39% (15)	4.46% (5)
Iranian (75)	93.33% (70)	0.00% (0)	6.67% (5)

A4. Study 4

A4.1 Publication of Study 4

Citation:
Roth, L.*, Huber, J. F.*, Kronenthaler, S., van Dijck, J., Cipora, K., Butz, M. V., & Nuerk, H.-C. (2025). Looks like SNARC spirit: Coexistence of short- and long-term associations between letters and space. *Quarterly Journal of Experimental Psychology, 78*(10), 2110–2132. https://doi.org/10.1177/17470218251324437

* shared first authorship

Copyright:
© 2025 Experimental Psychology Society (Open Access)
This article is published under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, provided the original authors and source are credited.

Check for updates

Original Article

QJEP

Quarterly Journal of Experimental Psychology
2025, Vol. 78(10) 2110–2132
© Experimental Psychology Society 2025

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17470218251324437
qjep.sagepub.com
Sage

Looks like SNARC spirit: Coexistence of short- and long-term associations between letters and space

Lilly Roth[1*], Julia F. Huber[1*], Sophia Kronenthaler[1], Jean-Philippe van Dijck[2,3], Krzysztof Cipora[4], Martin V. Butz[1,5] and Hans-Christoph Nuerk[1,6,7]

Abstract

Many studies have demonstrated spatial-numerical associations, but the debate about their origin is still ongoing. Some approaches consider cardinality representations in long-term memory, such as a Mental Number Line, while others suggest ordinality representations, for both numerical and non-numerical stimuli, originating in working or long-term memory. To investigate how long-term memory and working memory influence spatial associations and to disentangle the role of cardinality and ordinality, we ran three preregistered online experiments (N_{total} = 515). We assessed spatial response preferences for letters (which only convey ordinal but no cardinal information, in contrast to numbers) in a bimanual go/no-go consonant-vowel classification task. Experiment 1 ("no-go" trials: non-letter symbols) validated our setup. In Experiments 2 and 3, participants learned an ordinal letter sequence prior to the task, which they recalled afterwards. In Experiment 2, this sequence was merely maintained ("no-go" trials: non-letter symbols), whereas in Experiment 3, it needed to be retrieved during the task ("no-go" trials: letters outside the sequence). We replicated letter-space associations based on the alphabet stored in long-term memory (i.e., letters earlier/later in the alphabet associated with left/right, respectively) in all experiments. However, letter-space associations based on the working memory sequence (i.e., letters earlier/later in the sequence associated with left/right, respectively) were only detected in Experiment 3, where retrieval occurred during the task. Spatial short- and long-term associations of letters therefore seem to coexist. These findings support a hybrid model that incorporates both short- and long-term representations, which applies similarly to letters as to numbers.

Keywords

SNARC effect; letters; working memory; Ordinal Position Effect; Spatial-Alphabetical Association of Response Codes; SAARC effect

Received: 7 June 2024; revised: 23 September 2024; accepted: 29 October 2024

[1]Department of Psychology, University of Tübingen, Tübingen, Germany
[2]Department of Applied Psychology, Thomas More University of Applies Sciences, Antwerp, Belgium
[3]Department of Experimental Psychology, Ghent University, Ghent, Belgium
[4]Centre for Mathematical Cognition, Loughborough University, Loughborough, UK
[5]Department of Computer Science, University of Tübingen, Tübingen, Germany
[6]LEAD Graduate School & Research Network, University of Tübingen, Tübingen, Germany
[7]German Center for Mental Health (DZPG), Berlin, Germany

*Equal contribution, shared first-authorship

Corresponding author:
Hans-Christoph Nuerk, Department of Psychology, University of Tübingen, Schleichstraße 4, 72076 Tübingen, Germany.
Email: hc.nuerk@uni-tuebingen.de

Numbers play a fundamental role in many contexts in humans' everyday life. How humans mentally represent and process them is therefore a fundamental research question within the field of cognition. In 1993, Dehaene and colleagues observed what became widely known as the Spatial-Numerical Association of Response Codes (SNARC) effect. This association of small numbers with the left side and of large numbers with the right side has been demonstrated in many experiments, at least for Western cultures. The SNARC effect could be found in different tasks (see Wood et al., 2008 for a meta-analysis) and with different stimuli (e.g., digits, number words in visual or auditory modality or dice patterns; Nuerk et al., 2005). Most commonly, the effect is investigated with classification tasks like parity judgement (where participants classify numbers as even or odd; e.g., Dehaene et al., 1993) or magnitude classification (where participants classify a number as smaller or larger than a given reference number; e.g., Dehaene et al., 1990). Interestingly, the SNARC effect also occurs in tasks where the number magnitude is task-irrelevant, meaning it is not necessary to access the number magnitude information to solve the task correctly (such as in parity judgement, Dehaene et al., 1993, or font colour discrimination, Roth et al., 2025a). This has been interpreted as evidence for an automatic processing of number magnitude (Fias et al., 2001; Gevers et al., 2003).

Spatial associations of quantity and ordinality

Not only number magnitude but also non-numerical quantities (e.g., physical size or luminance; Cohen Kadosh et al., 2008; Fumarola et al., 2014; Ren et al., 2011) can be associated with space. These observations are accounted for by a Theory of Magnitude (ATOM; Bueti & Walsh, 2009; Walsh, 2003). Moreover, well-known ordinal sequences might also be related to space (e.g., letters of the alphabet; Gevers et al., 2003). According to Casasanto and Pitt (2019), referring to the ordinal position of objects in a sequence rather than to their quantity is sufficient to explain spatial mappings such as the SNARC effect. In contrast, Prpic et al. (2016) proposed that mechanisms underlying the SNARC effect depend on whether a *direct* task (i.e., magnitude- or order-relevant instructions, like magnitude classification) or an *indirect* task (i.e., magnitude- or order-irrelevant instructions, like parity judgement) is used (see also Prpic et al., 2021). This conclusion was derived from direct (bimanual note-value comparison) and indirect (unimanual note detection) tasks with musical note values, where order (usually starting from the whole note followed by the half note, the quarter note, the eight note, etc.) is opposed to magnitude (starting from the shortest duration and ending with the whole note). Thus, while mainly ordinal information drives the SNARC effect in direct tasks, magnitude information drives the SNARC effect in indirect tasks more than ordinal information does (see Cipora et al., 2020; Schroeder et al., 2017a, for integrative views).

Interestingly, spatial mappings have been observed both for arbitrary sequences trained within an experiment and stored in working memory (WM; e.g., van Dijck & Fias, 2011) as well as for well-known ordinal sequences stored in long-term memory (LTM; e.g., months and letters in Gevers et al., 2003; well-trained but arbitrary lists of objects in Previtali et al., 2010). The order of an overlearned sequence and the order of a sequence trained within an experiment can be pitted against each other to see which of them is decisive for spatial mapping (Koch et al., 2023). Importantly, using non-numerical stimuli with an overlearned order, one can distinguish between the spatial mapping of LTM and WM ordinality, while entirely excluding magnitude information. Given that previous results for direct tasks speak for the role of ordinality, we focus here on an indirect task. In this article, when we speak of *direct* or *indirect* tasks with non-numerical stimuli, we refer to whether participants are explicitly asked to process ordinality or not, independent of whether the long-term ordinality (i.e., canonical order stored in LTM, e.g., the alphabetical order for letters) or short-term ordinality (i.e., ordered sequence stored in WM) is meant (see differentiation in the literature in the online Supplementary Material C, Table C1).

Although Prpic et al. (2016) postulate that cardinality is more decisive than ordinality for the spatial mapping of numbers in indirect tasks, the spatial mapping of ordinality stored in LTM has been observed in indirect tasks with other stimuli (e.g., consonant-vowel judgement; Gevers et al., 2003, Exp. 2). A spatial mapping of ordinality has been also observed for WM content; however, to our understanding, despite being indirect, previous studies showing a WM-dependent mapping were making the WM demand salient to the participants (e.g., by requiring a constant retrieval of the WM content while performing the indirect task). Importantly, the spatial mapping of ordinality coming from WM may depend on the salience of this sequence stored in WM (i.e., how prominent the WM content is during the task to be performed), although the ordinality within this sequence might not be task-relevant or salient. Finding a spatial association in an indirect judgement task, where neither processing cardinality nor ordinality is required, can be seen as stronger evidence for the significant role of ordinality in spatial mental representations. To disentangle spatial associations consistent with ordinal information stored in WM vs. LTM, we investigated spatial associations of letters (which convey ordinal but not cardinal information, unlike numbers) in indirect tasks and manipulated how salient the retrieval of sequences stored in WM were during the task.

The role of WM and LTM ordinality for the SNARC and SNARC-like effects

The question arises as to how spatial associations of numbers or other kinds of stimuli with cardinality or ordinality properties can be explained. In their seminal article, Dehaene et al. (1993) described different explanations for the origins of spatial-numerical associations that are still part of the debate today: The SNARC effect is either based only on the representation of number magnitude / quantity (i.e., cardinality), or it is the result of an overlearned sequential structure of the stimuli (i.e., ordinality). Since Dehaene et al. (1993; Experiment 4) found no response preference for letters, they concluded that the SNARC effect arises from the magnitude representation, which can be imagined as a left-to-right oriented Mental Number Line (MNL; Moyer & Landauer, 1967; Restle, 1970). Note that Dehaene et al. (1993) only tested 10 participants in their consonant-vowel classification task (Experiment 4), so a true underlying effect might just have remained undetected due to poor statistical power for small effects. This is plausible, as the SNARC effect for Arabic digits in parity judgement has been shown not to be present in each individual and not to be stable over time within individuals (Roth et al., 2024). Crucially, Gevers et al. (2003, 2004) showed spatial response preferences for various non-numerical stimuli, namely letters, months, and days of the week, in both direct (e.g., order-relevant judgements on whether a letter comes before or after the letter O in the alphabetical order) and indirect (e.g., order-irrelevant judgements on whether the letter is a consonant or a vowel) task instructions. This effect has been called the Spatial-Alphabetical Association of Response Codes (SAARC) effect because letters early/late in the alphabet were associated with the left/right, respectively). Accordingly, the explanation of the SNARC effect being the result of the sequential structure of the stimuli (i.e., the inherent ordinal aspect of numbers) rather than an underlying spatially oriented MNL cannot be ruled out.

Van Dijck and Fias (2011) developed a theory that can explain spatial associations of numerical and non-numerical stimuli with ordinality, like letters or months (Gevers et al., 2003) or newly memorised arbitrary sequences of words or symbols (Previtali et al., 2010; van Opstal et al., 2009). Instead of assuming a stable spatial long-term representation of numbers in the form of a MNL, they postulated the involvement of WM in the emergence of spatial-numerical associations (i.e., the WM account; van Dijck et al., 2014; see Abrahamse et al., 2016, for later theoretical developments). Specifically, they assume that the SNARC effect is a result of the ordinal position of the elements stored in WM such that stimuli at the beginning of a sequence are associated with the left and stimuli at the end of a sequence are associated with the right (e.g., ordered sequences of numbers, but also of fruits and vegetables, as in van Dijck and Fias, 2011).

This WM account is supported by two empirical findings: First, WM seems to be involved in the emergence of the SNARC effect, as the SNARC effect disappeared under WM load depending on the task and the type of WM load. Namely, the SNARC effect disappears under verbal WM load in parity judgement (i.e., indirect task) and under spatial WM load in magnitude classification (i.e., direct task; Herrera et al., 2008; van Dijck et al., 2009). These findings demonstrate that WM capacity is required for the SNARC effect to arise, although the manifestation of this effect depends on the type of information held in WM (i.e., verbal or spatial). Second, the WM account is supported by the observation that the association of numbers and space is very flexible and varies depending on task demands (for an overview about situated influences on spatial-numerical associations, see Cipora et al., 2018). Some exemplary findings speaking in favour of the WM account are that the spatial associations depend only on the relative number magnitude within the used number range (Dehaene et al., 1993; Fias et al., 1996; but see Roth et al., 2025b), the impact of the number context (e.g., ruler vs. clock) imagined prior to the task (Bächtold et al., 1998; but see also Mingolo et al., 2024), or the location of numbers in a text (Fischer et al., 2010). These findings cannot be explained with a stable long-term association between space and numbers but rather suggest the involvement of WM. The spatial-numerical associations are assumed to be due to the position of the numbers (or any other stimuli) in WM and, thus, are rather spatial-positional associations (van Dijck & Fias, 2011).

To test their account, van Dijck and Fias (2011) used a thoughtful design, which combined the typical parity judgement task with a go/no-go paradigm, where the go/no–go classification relates to a previously learned sequence. Participants had to memorise a randomly generated sequence of five numbers between 1 and 10. Afterwards, they performed parity judgement, but only for numbers that were part of the memorised number sequence ("go" stimuli). Numbers that were not part of the memorised number sequence were not to be judged ("no-go" stimuli). After the task, it was checked whether participants had correctly memorised the sequence of numbers and only those blocks were included in subsequent analyses. Van Dijck and Fias (2011) observed that the spatial response preferences no longer relate to the number magnitude (i.e., no SNARC effect) but to the position of the numbers within the memorised number sequence. This phenomenon has been referred to as the Ordinal Position Effect (OPE; Ginsburg et al., 2014; which is the term we will use in the current manuscript) or as SPoARC (Spatial-Positional Association of Response Codes) effect (Guida & Lavielle-Guida, 2014).

In their recent opinion article, Abrahamse and van Dijck (2023) argue that ordinality eliciting the OPE refers to the serial position of an item in a so-called ranking-space relative to other items in that ranking-space. If no

specific manipulation is implemented, the set of the stimuli used in the study is encoded as such a sequence (i.e., numbers from 1 to 9 in the case of most SNARC studies). Importantly, when testing whether spatial response preferences are driven by magnitude or by ordinality, it is crucial to differentiate between ordinality based on short-term representations (i.e., newly learned arbitrary sequences stored in WM) or long-term representations (i.e., overlearned ordered sequences stored in LTM).

In Experiment 2, van Dijck and Fias (2011) were able to replicate the OPE using non-numerical stimuli. Importantly, the manipulation also worked with learned sequences of fruits and vegetables, which do not have any inherent order (as compared to letters, months, and days of the week). Ginsburg et al. (2014) replicated the results of van Dijck and Fias (2011, Exp. 1) using number stimuli, ruling out that the presence of the OPE and the absence of the SNARC effect was due to specific aspects of the design like the task (parity judgement or magnitude classification) or the response-to-key assignment (switch within or between participants).

Importantly, Ginsburg et al. (2014) found that the retrieval of the memorised sequence during the classification task (as in their Experiments 1 and 2 but not 3 and 4) is necessary to observe an OPE. In van Dijck and Fias' (2011) design the retrieval of the memorised sequence during the indirect classification task was ensured by the go/no-go design, as in the classification phase participants had to respond only to numbers that were part of the memorised sequence (i.e., high WM salience). When participants were instructed to respond to all numbers independently from the previously learned sequence (i.e., no WM salience), no OPE, but a SNARC effect was found (Ginsburg et al., 2014). In these conditions, participants memorised the numbers for a later recall and maintained them, whereas retrieval of the numbers during the direct or indirect classification tasks was not required. Interestingly, even the SNARC effect reflecting the spatialization of LTM ordinality or cardinality might depend on the sequence maintained in WM, although the sequence was irrelevant for the task: In the respond-to-all magnitude classification, a SNARC effect was only observed for the numbers that were part of the memorised sequence but not for other numbers. In line with the findings by Ginsburg et al. (2014), van Dijck et al. (2014) found an OPE for numbers in a Posner cueing task with (Experiment 2) but not without retrieval of the memorised sequence (Experiment 1). The OPE in the Posner cueing task with retrieval was replicated for letters (Experiment 3). Thus, the activation of the memorised sequence in WM by its retrieval seems to be a prerequisite for the OPE.

Storing a sequence in WM was only a secondary task in the experiments by van Dijck and Fias (2011) and Ginsburg et al. (2014). Nevertheless, the salience of WM was high in the judgement task because of the go/no-go design, which might have triggered an OPE. The literature overview in Table C1 (online Supplementary Material C) shows that the WM salience can be evaluated as high in all go/no-go designs that have so far been used. In this regard, the current study is unique, as it combines the instruction to store an ordered sequence in WM with a go/no-go design, while achieving different levels of WM salience, as described further below.

Spatial mapping of non-numerical ordered stimuli

In its original form, the WM account excluded the coexistence of the SNARC effect and the OPE; the SNARC effect should disappear or be reduced if a deviating order should be remembered in WM. However, as opposed to the findings by van Dijck and Fias (2011) and by Ginsburg et al. (2014), the coexistence of the SNARC and the OPE has been observed in later studies by Ginsburg and Gevers (2015) and by Huber et al. (2016). Moreover, ordinality has been demonstrated to influence spatial response preferences for numbers, even in cases where no ordinal number sequence is learned prior to the task. This was illustrated by Koch et al. (2023). In a parity judgement setup, they disentangled the inherent ordinality of numbers from their magnitude by using number sets such as 1, 2, 3, and 8 and testing whether ordinality or magnitude determines the pattern of response times (RT) better. Typically, the SNARC effect and the OPE are quantified by the slopes obtained by regressing, for each participant separately, the dRTs (i.e., the difference in the average RT between right- and left-hand responses for each magnitude or position) to the magnitudes or positions. In this example, there are five magnitude units between 3 and 8, but these are the third and fourth numbers when ordering the set by increasing magnitude so that there is only one ordinality unit between 3 and 8. Koch et al. (2023) investigated whether both numbers or positions explained a unique proportion of variance in dRTs, and found influences by both inherent ordinality and magnitude for the spatialization of numbers. Ginsburg and Gevers (2015) assume that the SNARC effect and the OPE activate different representations and can therefore coexist, further suggesting that the MNL and the WM account are not mutually exclusive. More precisely, the SNARC effect reflects the spatial representation of the canonical order of numbers being stored in LTM, whereas the OPE reflects the spatial representation of either an overlearned order in LTM (e.g., the alphabet in the case of letters) or a newly learned order being stored in WM. Abrahamse et al. (2016) claim that a template for a spatial mental representation is most likely stored in LTM, but the template is then retrieved from LTM and used in WM. According to their view, the items under question are filled into this template only in WM, no matter whether a sequence is already familiar (i.e., stored in LTM and

retrieved to WM) or only newly learned (i.e., only stored in WM, but not in LTM). In contrast, Guida and Campitelli (2019) claim in their expertise account that spatial information can be stored together with orders in LTM and that the association between orders and space is not always built in WM. Specifically, in their view, spatialization already happens in LTM for canonical orders (such as of numbers or letters), and only takes place in WM for newly learned orders, for which no spatial information exists in LTM. It is not clear, however, how this account can explain the disappearance of the SNARC effect when WM resources are depleted (e.g., Ginsburg et al., 2014; van Dijck et al., 2009). The simultaneous observation of the SNARC effect and the OPE led to the development of a new hybrid explanation: Not a single underlying representation determines spatial-numerical associations, but instead an interplay of two or more components, such as LTM representations and WM contents (van Dijck et al., 2015). In line with this, differential effects of anodal tDCS (transcranial Direct Current Stimulation) have been found for different stimulus types (i.e., number words, weekdays, and months), suggesting that multiple codes (e.g., verbal markedness, spatial verbal instruction, sequential WM, and visuospatial stimulation) are responsible for spatial associations of ordinal stimuli (Schroeder et al., 2017b).

As mentioned earlier, previous research has been able to show spatial associations not only for numbers but also for non-numerical stimuli. The current study aimed at answering the question of whether the WM account also holds for other ordered stimuli that do not convey any magnitude information (i.e., metathetic dimensions, for which ATOM is not claimed to hold), such as letters. Similar to numbers, letters have a fixed inherent order (i.e., we count "one, two, three,...," and we recite the alphabet "A, B, C,..."). If we compare numbers with letters or other familiar sequences such as the alphabet, numbers are special. The order of numbers is based on their cardinality (e.g., the number 3 corresponds to three elements and the number 4 to four elements), and they have a quantitative meaning (e.g., 4 is larger than 3). This is not the case for letters (e.g., B is not more or less than A): While numbers convey both magnitude and ordinality, letters possess only ordinality (e.g., B comes after A in the alphabet) without magnitude. As shown by Zorzi et al. (2006) in mental interval bisection tasks, left-spatial-neglect patients' performance was influenced by the interval length for number stimuli (i.e., bias to the left in small intervals like 1–3, but bias to the right in large intervals like 1–9) but not for letter stimuli (e.g., no evidence for a systematic performance difference between L–N and L–T). The authors thus concluded that the mental representations of letters and numbers are qualitatively different. In conclusion, it remains uncertain what spatial mental representations of non-numerical metrics, characterised by ordinal but lacking cardinal information, entail and whether they are governed by shared processes with spatial associations of numbers.

The current study and hypotheses

To find out whether spatial associations for ordinal sequences depend on the salience of WM demands, we investigated letter-space associations while manipulating the degree to which WM is involved in the task. Crucially, by using letters, we were able to disentangle ordinality from magnitude in three experiments. We adapted the standard WM paradigm of van Dijck and Fias (2011), taking into account findings by Ginsburg et al. (2014) that the retrieval of the memorised sequence during the classification task is essential to observe an OPE. Experiment 1 did not include any encoding phase (i.e., no WM salience) and served as manipulation check to see whether a task set up with a go/no-go design only influences spatial associations (i.e., *control condition*). In two further experiments, we manipulated the need to retrieve the memorised letter sequence during the classification task by using different "no-go" stimuli. In Experiment 2, we used non-letter symbols like ‡, △, ≡, and ∇ as "no-go" stimuli, which could easily be identified as a "no-go" trial. The memorised sequence had to be maintained during the classification task but needed to be recalled only later (i.e., *maintenance condition*), creating intermediate WM salience. In Experiment 3, we used letters that were not part of the letter sequence memorised in the encoding phase (i.e., *retrieval condition*) as "no-go" trials. Thus, it was necessary to retrieve the memorised sequence to identify a letter as a "go" trial (i.e., when it was part of the memorised sequence) or as a "no-go" trial (i.e., when it was not part of the memorised sequence), creating high WM salience.

According to Ginsburg et al. (2014), an OPE is only elicited if the memorised ordinal sequence needs to be retrieved from WM during the judgement task. In contrast, only spatial associations based on LTM can be observed if the memorised sequence is merely maintained in WM without being retrieved. Hence, if the same mechanisms apply to letters as to numbers, no OPE but a SAARC effect should be found in the maintenance condition (Experiment 2). At the same time, an OPE but no SAARC effect should be found in the retrieval condition (Experiment 3).

The present study was preregistered (https://osf.io/945uq), and all datasets are publicly available on the Open Science Framework, together with the corresponding R scripts for the analysis (https://osf.io/v84m5/). Note that in the preregistration, we wrote that we wanted to compare the strength of both effects with one another; however, upon closer inspection, we realised that it is not the difference between the two effects that is meaningful for the underlying mechanism. Instead, what matters is solely the presence of an OPE as a marker for WM involvement or of a SAARC as a marker for LTM involvement.

We therefore adapted our hypotheses as described above and our analyses as described below.

General method

Overview and design

In three online experiments with a between-subjects design, which differed in the salience of the WM instruction, we used a go/no-go paradigm that closely followed the study by van Dijck and Fias (2011) and subsequent studies (Ginsburg et al., 2014; Ginsburg & Gevers, 2015). The main differences were (i) that stimuli were letters instead of numbers to rule out cardinality and (ii) that the used indirect task was a consonant-vowel classification instead of a parity judgement. Furthermore, (iii) each participant had to remember only one sequence across the entire experiment in the initial encoding phase. In contrast, previous studies consisted of several experimental blocks each including (1) an encoding phase, (2) a classification phase, and (3) a control phase, such that there was one new sequence per block (number of blocks per study: 20 respond-to-all or 20 go/no-go in Ginsburg et al., 2014; 20 respond-to-all mixed with 20 go/no-go in Ginsburg & Gevers, 2015; 54 go/no-go in Huber et al., 2016; 20 go/no-go in van Dijck and Fias, 2011). Importantly, the salience of WM demands was systematically increased from Experiment 1 (no WM salience: no ordinal sequence maintained) over Experiment 2 (intermediate WM salience: ordinal sequence maintained and retrieved after the task) to Experiment 3 (high WM salience: ordinal sequence maintained and retrieved during the task), as illustrated in Table 1.

Sample size

We conducted an a priori sample-size determination analysis in JAMOVI (The Jamovi Project, 2023) to determine sample size (see our preregistration: https://osf.io/945uq). To obtain a power of 0.8 for detecting the minimally relevant effect size of Cohen's $d = 0.25$ with the Bonferroni-corrected significance level of $\alpha = .05 / 3 = .017$ in one-tailed independent-samples *t*-tests to compare the SAARC effect and the OPE between experiments, a sample size of 282 participants per experiment (thus a total $N = 846$) would have been required. Between July and December 2021, the target sample size was not achieved, despite intensive recruitment efforts (via round mails at the University of Tübingen, Facebook, Instagram, Twitter, Reddit, and personal acquaintances), and despite adding an English version, which might be due to the COVID-19 pandemic. We stopped data collection with a total of 515 participations (German version: 474, English version: 41).

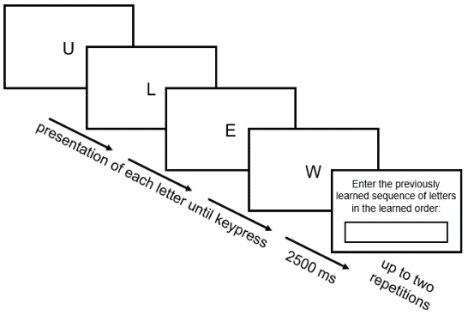
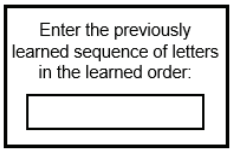
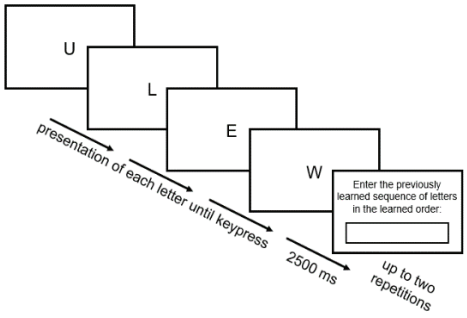
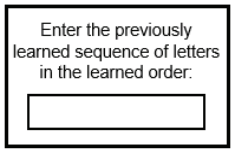
Importantly, instead of running three independent-samples *t*-tests as stated in the preregistration, we ran five one-sample *t*-tests, as they fit our research questions better. Therefore, the preregistered sample-size determination analysis was invalid for the analyses that we eventually decided to apply, and we conducted a power-determination analysis (as described by Giner-Sorolla et al., 2024, see R scripts at https://osf.io/v84m5/). The reached sample sizes prior to exclusions due to data preprocessing were 178, 169, and 168 for Experiments 1, 2, and 3, respectively; 149, 105, and 107 datasets remained for the analysis after preprocessing. The determined power levels for detecting the minimally relevant effect size of Cohen's $d = 0.25$ with the Bonferroni-corrected significance level of $\alpha = .05 / 5 = .01$, and the size of these final samples were poor, namely 0.43, 0.30, and 0.31. In a subsequent effect-size sensitivity analysis (as described by Giner-Sorolla et al., 2024), we found out that with a power level of 0.80, the applied statistical tests would be able to detect true underlying effects with $d = 0.37$, $d = 0.44$, and $d = 0.44$ with the final sample sizes. Note that the final samples are still more than four times larger than in the original studies by van Dijck and Fias (2011) and Ginsburg et al. (2014) on which our work is based and who tested between 21 and 39 participants per experiment, and that $d = 0.4$ is reasonable as the smallest effect size of interest in psychology according to Brysbaert (2019). All participants in the present study were adults with a minimum age of 18 years. As compensation, participants could either receive course credits or participate in a lottery, with the opportunity to win a voucher worth €50, with odds of one in seven.

Materials and stimuli

The three online experiments were programmed in jsPsych (de Leeuw, 2015). We hosted the experiments on SoSci Survey (Leiner, 2021), to assign participants randomly to the experiments and obtain their email addresses for the lottery, and on Pavlovia (Pavlovia Surveys (Open Science Tools), 2021), to run the experiment with the tasks and questionnaires on demographic and control variables.

To be able to compare results across the three experiments, we used the same stimulus sets in each experiment, with two sets being put together of the same four letters whose order was changed (see Table 2).[1] The letter sets were composed of two vowels and two consonants, with one vowel and one consonant on the left and right sides of the keyboard, respectively (see Figure 1). In addition, the letters were selected to cover a wide range of the alphabet, to be distributed as evenly as possible over the alphabet, and to not form words or familiar abbreviations (e.g., company names). The resulting letter sequences were [O J R A], [R A O J], [L E W U], and [U L E W] (see the online Supplementary Material A for a detailed description of the letter selection procedure).

Table 1. Scheme of the experimental procedure with ascending salience of WM demands from Experiment 1 to Experiment 3. Consonant-vowel classification was used in all three experiments. Experiments 1 and 2 only differed regarding the encoding and control phases, which were absent in Experiment 1 (i.e., without any WM task) and present in Experiments 2 and 3 (i.e., including a WM task). Experiments 2 and 3 only differed regarding the "no-go" stimuli, which were symbols in Experiments 1 and 2 (i.e., no retrieval from WM required for the distinction between "go" and "no-go" stimuli) and letters in Experiment 3 (i.e., retrieval from WM required).

1	none	'Go' stimulus: U 'No-go' stimulus: ∆	none
2		'Go' stimulus: U 'No-go' stimulus: ∆	
3		'Go' stimulus: U 'No-go' stimulus: A	

Each participant was assigned to only one stimulus set (out of two in Experiment 1, where the order of the letters did not matter, and out of four in Experiments 2 and 3, where the order of the letters did matter), which were counterbalanced between participants. The "go" stimuli in the classification phase consist of four different letters (in Experiments 2 and 3, the "go" stimuli were the four letters learned in the encoding phase). The "no-go" stimuli differed between the experiments: In Experiments 1 and 2, we used the symbols ‡, △, ≡, and ∇. In Experiment 3, we used the letters of the respective other stimulus set as "no-go" stimuli (e.g., when the stimulus set was [O J R A] or [R

Table 2. Overview of the memorised letter sequences as well as "go" and "no-go" stimulus sets, separately for each experiment, with each row corresponding to one between-subjects condition participants were randomly assigned to (note that within each of conditions, participants were randomly assigned to one of two response-to-key assignment orders: Participants either responded to vowels with D and to consonants with K in the first half and to consonants with D and to vowels with K in the second half, or vice versa).

Experiment	(1) encoding phase, (3) control phase	(2) classification phase	
	Memorised sequence	'Go' stimulus set	'No-go' stimulus set
Exp. 1	none	A, J, O, R	‡, △, ≡, ∇
	none	E, L, U, W	‡, △, ≡, ∇
Exp. 2	[O J R A]	A, J, O, R	‡, △, ≡, ∇
	[R A O J]	A, J, O, R	‡, △, ≡, ∇
	[L E W U]	E, L, U, W	‡, △, ≡, ∇
	[U L E W]	E, L, U, W	‡, △, ≡, ∇
Exp. 3	[O J R A]	A, J, O, R	E, L, U, W
	[R A O J]	A, J, O, R	E, L, U, W
	[L E W U]	E, L, U, W	A, J, O, R
	[U L E W]	E, L, U, W	A, J, O, R

A O J], the letters E, L, U, and W were used as "no-go" stimuli, see Table 2). Importantly, the used letters did not differ between experiments (see Table 2), so that differences between experiments cannot be attributed to stimuli used in the consonant-vowel classification task.

Procedure

This study was approved by the Ethics Committee of the University of Tübingen's Department of Psychology. In the beginning, participants were informed about the study, eligibility requirements (i.e., participation with a desktop, laptop, or netbook), and data protection. After their informed consent, they were randomly assigned to one of the three experiments, and subsequently to one of two "go" stimulus sets (in Experiment 1) or to one of four-letter sequences to be learned (in Experiments 2 and 3), as summarised in Table 2.

1. *Encoding phase*: In Experiments 2 and 3, participants were instructed to memorise a sequence consisting of four letters. The letters were presented one after another in the centre of the screen. Participants could move from one letter to the next via keypress so that they determined the presentation speed. After 2500 ms, participants should type the letter sequence. In case of a mistake, the letter sequence was presented again, with a maximum of three attempts.
2. *Classification phase*: In the main part of all three experiments, participants were instructed to decide via key press (D or K) whether a presented letter was a consonant or a vowel. In one half of the experiment, participants were instructed to press the left response key (D) when they saw a consonant and the right response key (K) when they saw a vowel. In the other half, the response-to-key assignment was reversed. Response-to-key assignment order was counterbalanced between participants. In Experiment 1, participants were instructed to respond to letters ("go" stimuli) but not to symbols ("no-go" stimuli, no WM salience). Participants were instructed to respond only to memorised letters ("go" stimuli) but not to symbols ("no-go" stimuli) in Experiment 2 (intermediate WM salience) or to letters that were not part of the learned letter sequence ("no-go" stimuli) in

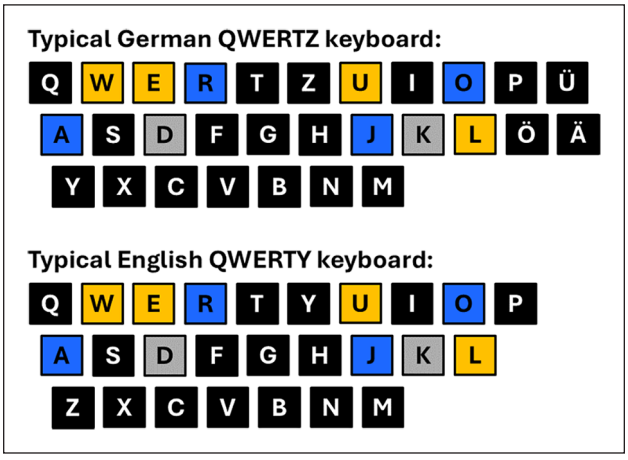


Figure 1. Letter positions on a typical German QWERTZ keyboard (upper panel) and on a typical English QWERTY keyboard (lower panel), with yellow/blue colour-coding for the used letter sets E, L, U, W and A, J, O, R, respectively, and with the response keys coloured in grey.

Experiment 3 (high WM salience). Each trial started with the presentation of a fixation cross for 500 ms, followed by the presentation of the stimulus until a response key was pressed or after a maximum of 1500 ms. During each inter-trial interval, a blank screen was presented for 100 ms. Each stimulus was repeated 20 times per response-to-key assignment, resulting in a total of 320 trials (8 stimuli x 2 assignments x 20 repetitions), that is 160 "go" trials and 160 "no-go" trials. The order of stimulus presentation was pseudorandomized, so that the same stimulus was never repeated in consecutive trials. After every 50 trials, there was an opportunity to take a short break. At the beginning of the respective experiment and after changing the response-to-key assignment, 16 practice trials were presented, with each stimulus (four "go" stimuli and four "no-go" stimuli) being presented twice. In practice trials, accuracy feedback was displayed for 1000 ms. If fewer than 75% of the stimuli were answered correctly, the practice trials were repeated.

3. *Control phase*: In Experiments 2 and 3, participants should type the letter sequence learned beforehand in the encoding phase (open response format). As in the encoding phase, the letters could be entered up to three times.

After these phases, a questionnaire was presented, asking participants about:

- memory strategy (open response format, e.g., whether they had written down the letter sequence to be memorised)
- age and gender (*male*, *female*, *diverse*)
- mother tongue and language skills (last grade: 0—*very poor* to 10—*very good*; diagnosed reading/spelling disability or dyslexia: *yes*, *no*)
- knowledge of a language with a reading/writing direction from right to left (*yes*, *no*)
- handedness (*right-handed*, *left-handed*, or *ambidextrous*), starting hand when counting with fingers (*left*, *right*), and stability of this starting hand (*always*, *usually*, *sometimes*)
- alphabetical knowledge (RT and correctness of the answer to four questions about the position of the letters in the alphabet: *Does the letter "K" come before "M" in the alphabet? Does the letter "T" come before "L" in the alphabet? Does the letter "R" come before "O" in the alphabet? Does the letter "J" come before "Q" in the alphabet?*)
- data quality (how noisy the environment was during participation: *extremely quiet*, *very quiet*, *a little quiet*, *a little noisy*, *very noisy*, *extremely noisy*; whether participants would use their data if they were the experimenter: *yes*, *not all*, *no*)

Data preprocessing

All data preprocessing steps, statistical power considerations, and data analyses were run in the statistical computing software R (R Core Team, 2023), using the R packages *clinfun* (Seshan & Whiting, 2023), *dplyr* (Wickham, François, et al., 2023), *ggplot2* (Wickham et al., 2024), *ggpubr* (Kassambara, 2023), *plyr* (Wickham, 2023a), *pwr* (Champely et al., 2020), *stringr* (Wickham, 2023b) *tidyr* (Wickham, Vaughan, et al., 2023), and *tidyverse* (Wickham, 2023c). The analysis R scripts as well as the anonymized data are available on the Open Science Framework at https://osf.io/v84m5/.

Our analysis was based on RT data (i.e., time from the stimulus onset until a key press) of the correct responses in the "go" trials. Exclusion of outliers and filtering of the data were performed in line with the preregistration (https://osf.io/945uq) and analogous to the procedure of Cipora et al. (2019). Specifically, we excluded entire datasets from:

- incomplete participations (i.e., participants quit before the end)
- participants who used other devices than a desktop, laptop, or netbook
- participants who performed the task in a very or extremely noisy environment
- participants who declared that they would not use their data or not all of it if they were the experimenter
- participants who correctly refrained from responding in fewer than 80% of all "no-go" trials
- participants who could not reproduce the learned letter sequence correctly at their first try in the control phases in Experiments 2 or 3 (note that we did not specify in the preregistration whether participants who could reproduce the sequence correctly at their second or third try would also be excluded, but since memorising the exact letter sequence correctly is a prerequisite for measuring the OPE, we decided to be strict here)
- participants who indicated that their memory strategy had been to write down the letter sequence to be memorised in the encoding phase (note that this criterion was not preregistered, but having the correct letter sequence in WM rather than using external aids is crucial for the OPE)

Afterwards, we removed all incorrectly responded trials and all trials with RTs below 250 ms (as it was done by

Cipora et al., 2019 and van Dijck & Fias, 2011). Then, as preregistered, a sequential trimming procedure was conducted within each participant, sequentially removing all RTs beyond ± 3 *SD* from the individual mean RT. Finally, full datasets of participants with fewer than 70% of valid trials left per letter in each response-to-key assignment were excluded.

Data analysis

The data analysis was based on the repeated-measures regression procedure proposed by Lorch and Myers (1990) and first applied to the SNARC effect by Fias et al. (1996). Accordingly, we first calculated mean RTs for each participant and each letter separately for both response sides. Then, mean RT differences (dRTs) were computed for each letter by subtracting mean RTs given with the left hand from mean RTs given with the right hand. Negative dRTs indicate faster responses with the right hand, while positive dRTs indicate faster responses with the left hand. Subsequently, linear regressions were performed for each participant *i*, with either (1) the position of the letters in the alphabet (e.g., 15, 10, 18, 1 for set [O J R A]) or (2) the position of the letters within the memorised letter sequence (i.e., 1, 2, 3, 4) as the predictor variable and dRTs as outcome variable:

$$\text{(1) SAARC: } dRT_i = \beta_0 + \beta_1 * position_{(alphabet)i} + \varepsilon_i$$

$$\text{(2) OPE: } dRT_i = \beta_2 + \beta_3 * position_{(sequence)i} + \varepsilon_i$$

Both regression models were fit for participants in Experiments 2 and 3, whereas only the Model (1) was fit for participants in Experiment 1, as no ordinal letter sequence was learned in this control condition. The unstandardized regression slopes β_1 and β_3 indicate the change in right- over left-hand advantage in milliseconds per increase of one position unit. The unstandardized slopes β_1 resulting from Model (1) indicate the SAARC effect, while the unstandardized slopes β_3 from Model (2) indicate the OPE. We standardised and Fisher-z-transformed the regression slopes within each experiment, such that they can be interpreted as *effect sizes*. The main dependent variable was standardised slopes, which are unaffected by individual RTs and individual variance in RTs. The presence of a SAARC effect and an OPE at the group level was tested by comparing these slopes to zero using a one-tailed, one-sample *t*-test (as in previous literature, e.g., Ginsburg et al., 2014; Ginsburg & Gevers, 2015; van Dijck & Fias, 2011). We accounted for multiple hypothesis tests via the Bonferroni-Holm correction,[2] starting with a significance level of $\alpha = .05 / 5 = .01$ for the test with the lowest corresponding *p* value. Note that we preregistered to run three tests because we originally planned to compare the SAARC effect and the OPE within each experiment. However, because quantitatively comparing slopes for two qualitatively different effects is theoretically not meaningful, we decided to test the SAARC effect and the OPE separately within each experiment, resulting in five tests. For data exclusions at the participant level and for a summary of the main results for all three experiments, see Tables 3 and 4, respectively.

Experiment 1

Gevers et al. (2003) showed spatial response preferences for letters based on their ordinal position in the alphabet: Letters at the beginning of the alphabet were responded to faster with the left than with the right hand, while the reverse response pattern was observed for letters at the end of the alphabet (see their Experiment 2, order-irrelevant task). In Experiment 1, in addition to letters as "go" stimuli, symbols were presented as "no-go" stimuli. This control condition served as a manipulation check and makes it possible to integrate our findings into the existing literature on response preferences for letters, as this is the first study on letter-space associations using a go/no-go design. We expected faster left-hand responses for letters at the beginning of the alphabet and faster right-hand responses for letters at the end of the alphabet (i.e., a SAARC effect), replicating the results by Gevers et al. (2003). As no letter sequence had to be learned in this experiment, WM was not salient and no OPE could be computed for this condition.

Method

Participants. In all, 178 individuals participated in this online experiment. After exclusions (see Table 3), datasets from 149 participants (German version: 136, English version: 13) remained for the analysis. The average age of these participants was 26.75 years (*SD* = 10.20, range: 18 to 72). Ninety-nine participants identified as female and 50 as male; 130 participants reported being right-handed, 16 left-handed, and three ambidextrous; and 130 participants indicated German as their native language (other native languages were English, Chinese, Turkish, Spanish, Bengali, Portuguese, Taiwanese, Hungarian, French, and Hindi; none of them with reading/writing direction from right to left).

Stimuli, design, and procedure. Participants completed a consonant-vowel classification task in which "no-go" stimuli were ‡, △, ≡, and ∇. They were assigned to one of four conditions in a 2 ("go" stimuli: A, J, O, R vs. E, L, U, W) × 2 (response-to-key order: start with vowel-left and consonant-right vs. start with vowel-right and consonant-left) design, as illustrated in Table 2.

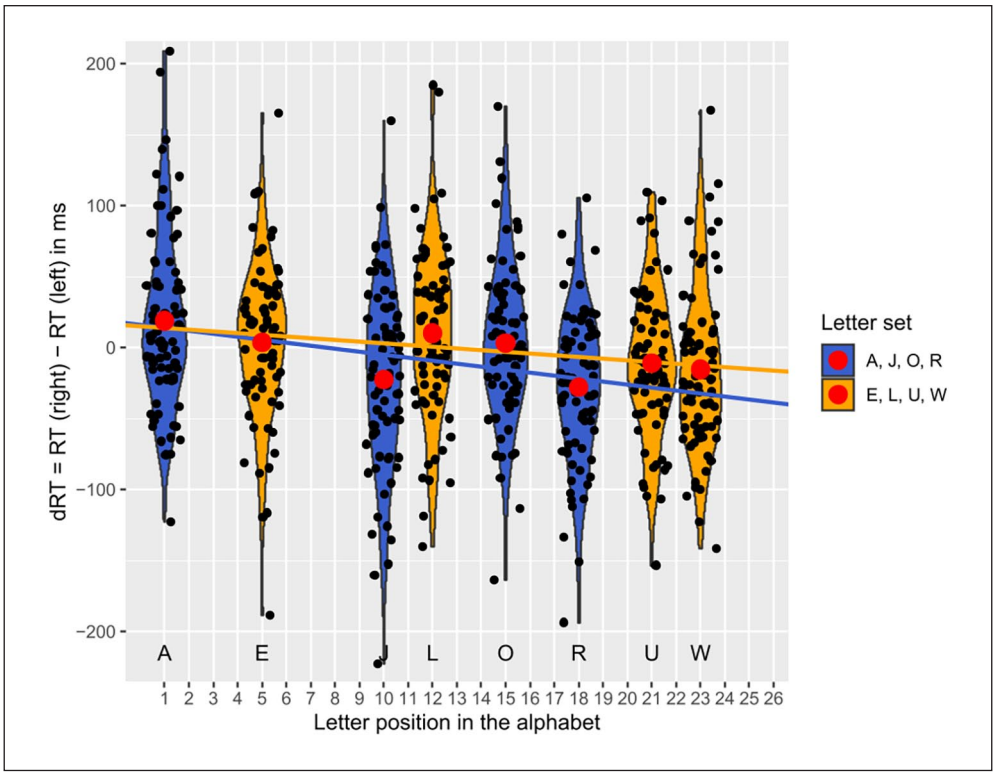


Figure 2. Differences in reaction times (dRT) with the right and left hand depending on the position of the letter in the alphabet (SAARC effect) in Experiment 1 (manipulation check), with black dots being jittered mean dRTs per participant for the respective position surrounded by violins, red dots being mean dRTs averaged over participants, and regression lines reflecting the SAARC slopes averaged over participants (between-subjects assignments are colour-coded in yellow/blue for the letter sets E, L, U, W and A, J, O, R, respectively).

Results

After applying exclusion criteria at the participant level (except for two: minimum of 80% valid "no-go" trials and minimum of 70% valid "go" trials per cell), 6.96% of the trials were excluded from the analysis (first 5.53% due to incorrect responses, another 0.02% because of anticipations with RTs below 250 ms, and another 1.41% within the sequential trimming procedure). The mean RT was 584.7 ms ($SD = 134.5$ ms). Mean RTs per letter and mean accuracies per letter set can be found in the online Supplementary Material B (Tables B1 and B2).

The regression analysis revealed a significant negative standardised slope for the letter position in the alphabet, $\beta_1 = -0.42$, $t(148) = -5.54$, $p < .001$ (using $\alpha = .0125$). Thus, as expected, a significant SAARC effect was found, indicating an increase of right-hand advantage with later positions of the letter in the alphabet, see Figure 2.

Discussion

These results are a replication of Gevers et al.'s (2003) and the observed SAARC effect confirms the spatial response preferences for letters according to their order in the alphabet (i.e., according to their ordinal representation in LTM). To conclude, the consonant-vowel classification task including non-letter "no-go" trials can be used to measure spatial associations.

Experiment 2

According to Ginsburg et al. (2014), in the case of numerical stimuli, the observation of the OPE depends on the retrieval of the learned sequence. If the sequence was only maintained for later retrieval, no OPE was found; instead, a regular SNARC effect emerged. In this experiment, we investigate whether this pattern can also be shown for non-numeric ordinal sequences like letters. Analogous to the results of Ginsburg et al. (2014), we assume that, if the learned sequence was merely to be maintained in WM for later recall, and the WM salience in the judgement task was thus at best intermediate, no OPE but a SAARC effect should be found.

Method

Participants. The sample consisted of 169 participants. After exclusions (see Table 3), datasets of 105 participants

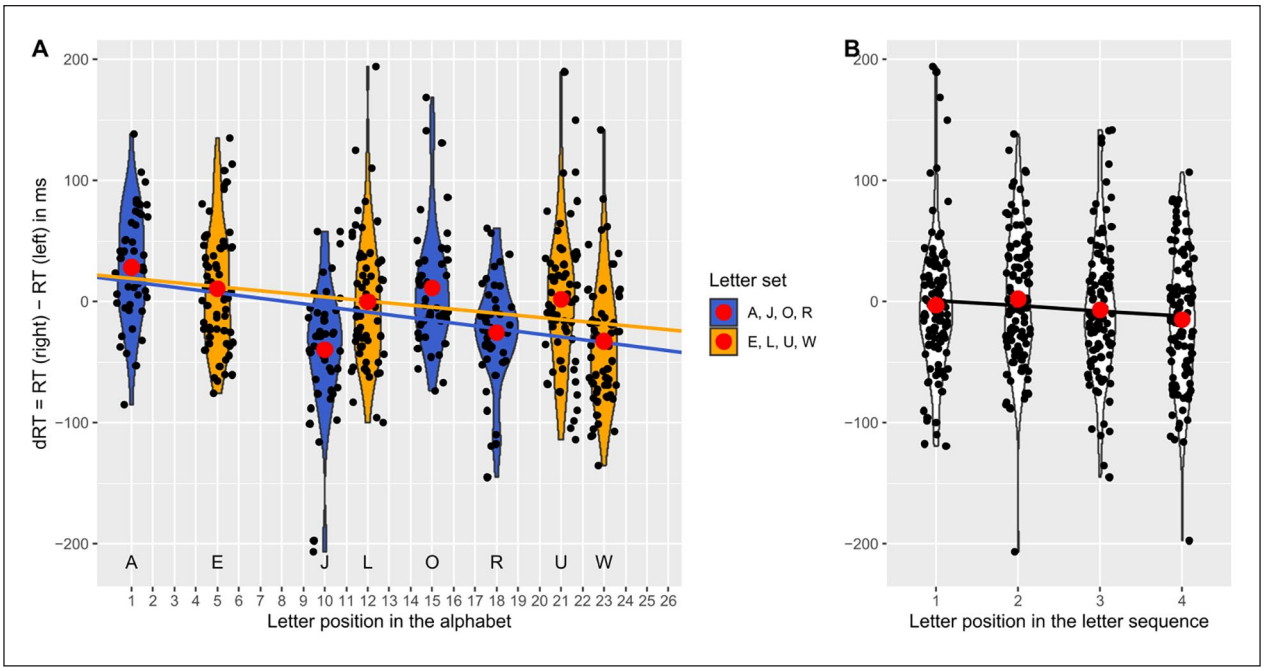


Figure 3. Differences in reaction times (dRT) with the right and left hand depending on the position of the letter in the alphabet (SAARC effect, Panel A) or in the memorised sequence (OPE, Panel B) in Experiment 2 (maintenance condition), with black dots being jittered mean dRTs per participant for the respective position surrounded by violins, red dots being mean dRTs averaged over participants, and regression lines reflecting the SAARC slopes averaged over participants in Panel A (between-subjects assignments are colour-coded in yellow/blue for the letter sets E, L, U, W and A, J, O, R, respectively) and the OPE slopes in Panel B.

(German version: 97, English version: 8) with an average age of 25.79 years (SD = 10.71, range: 18 to 71) were analysed. Of these participants, 71 identified as female, 33 as male, and one as non-binary. 98 participants reported to be right-handed, five to be left-handed, and two to be ambidextrous. The native language was German for 90 participants (other native languages were English, German, Luxembourgish, Turkish, Russian, Chinese, Arabic, Italian, Greek, Hindi, and Spanish). Only one participant reported a native language with reading/writing directions from right to left.[3]

Stimuli, design, and procedure. As in Experiment 1, participants completed a consonant-vowel classification task in which "no-go" stimuli were ‡, △, ≡, and ∇. They were assigned to one of eight conditions in a 4 ("go" stimuli and letter sequence: [O J R A] vs. [R A O J] vs. [L E W U] vs. [U L E W]) × 2 (response-to-key order: start with vowel-left and consonant-right vs. start with vowel-right and consonant-left) design, as illustrated in Table 2.

Results

After applying exclusion criteria at the participant level (except for two: minimum of 80% valid "no-go" trials and minimum of 70% valid "go" trials per cell), 5.79% of the trials were excluded from the analysis (first 4.24% due to incorrect responses, none because of anticipations with RTs below 250 ms, but another 1.55% within the sequential trimming procedure). The mean RT in the classification task was 563.5 ms (SD = 124.5 ms). Mean RTs per letter and mean accuracies per letter set can be found in the online Supplementary Material B (Tables B1 and B2).

The first regression analysis revealed a significant negative standardised slope for the letter position in the alphabet, $\beta_1 = -0.40$, $t(104) = -6.00$, $p < .001$ (using $\alpha = .0100$). Thus, as expected, and as in Experiment 1, a significant SAARC effect was found, indicating an increase of right-hand advantage with later positions of the letter in the alphabet (see Figure 3, Panel A). In contrast, the second regression analysis did not reveal a significant negative standardised slope for the letter position in the learned sequence, $\beta_3 = -0.08$, $t(104) = -1.08$, $p = .142$ (using $\alpha = .0500$). Thus, as expected, no significant OPE was found (see Figure 3, Panel B).

Discussion

Our results extend the results of Ginsburg et al. (2014) to non-numerical ordinal stimuli like letters. Like Ginsburg et al. (2014), we found no OPE when the memorised sequence was merely maintained for a later recall. Instead, in line with their finding of a SNARC effect, we observed a SAARC effect (i.e., a spatial mapping according to the ordinal representation of letters in LTM).

Experiment 3

As explained above, Ginsburg et al. (2014) found that the OPE for numerical stimuli depends on the retrieval of the

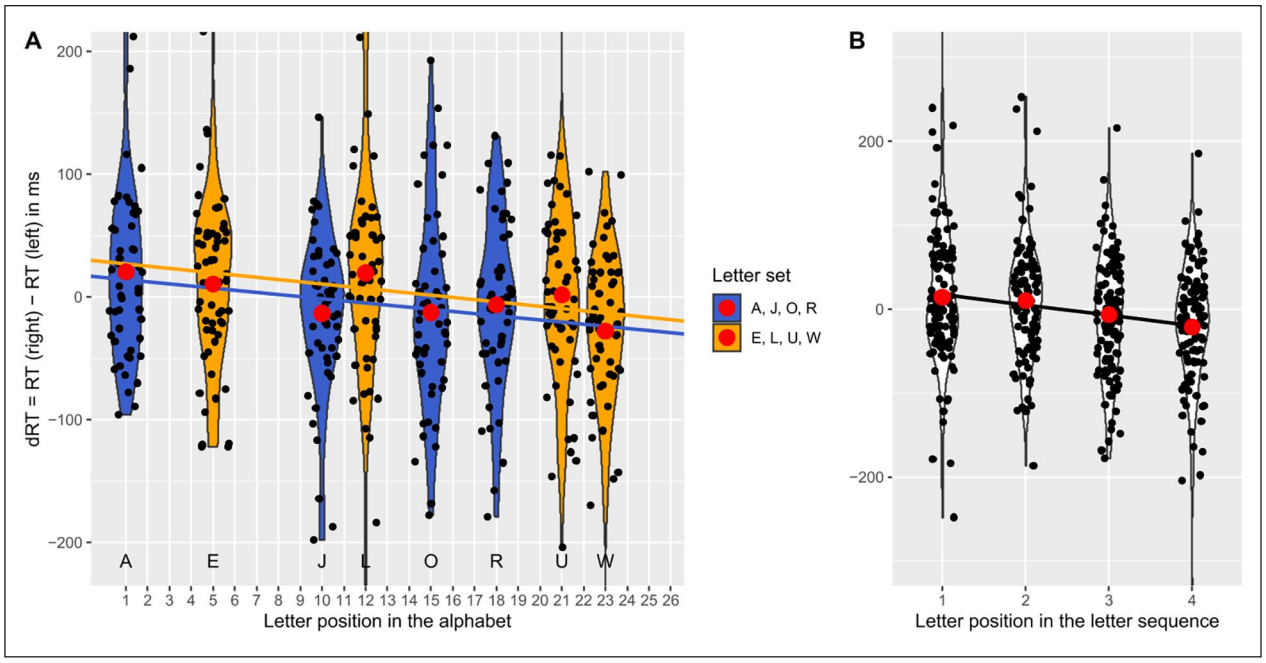


Figure 4. Differences in reaction times (dRT) with the right and left hand depending on the position of the letter in the alphabet (SAARC effect, Panel A) or in the memorised sequence (OPE, Panel B) in Experiment 3 (retrieval condition), with black dots being jittered mean dRTs per participant for the respective position surrounded by violins, red dots being mean dRTs averaged over participants, and regression lines reflecting the SAARC slopes averaged over participants in Panel A (between-subjects assignments are colour-coded in yellow/blue for the letter sets E, L, U, W and A, J, O, R, respectively) and the OPE slopes in Panel B.

learned sequence during the classification task. To investigate whether this is also true for non-numerical ordinal stimuli like letters, we used letters deviating from the memorised letter sequence as "no-go" stimuli during a consonant-vowel classification task. Using other letters as "no-go" stimuli requires participants to retrieve the memorised letter set during the classification task itself (i.e., retrieval condition), creating high WM salience. We expected the same mechanism to apply to letters as to numbers, such that an OPE but no SAARC effect would be found in this experiment.

Method

Participants. In all, 168 individuals participated in this experiment. After exclusions (see Table 3), the final sample for the data analysis consisted of 107 participants (German version: 100, English version: 7) with an average age of 25.43 years (SD = 7.31, range: 18 to 64). 75 of them identified as female, 30 as male, and two as non-binary. Furthermore, 96 participants indicated being right-handed, nine left-handed, and two ambidextrous. 96 participants reported German as their native language (other reported native languages were English, Turkish, Kurdish, Bulgarian, Polish, Greek, Bosnian, Spanish, Ukrainian, Persian, and Hindi). Only one participant indicated a native language with reading and writing directions from right to left.[4]

Stimuli, design, and procedure. As opposed to the previous two experiments, the "no-go" stimuli were letters instead of symbols. Participants were assigned to one of eight conditions in a 4 ("go" stimuli and letter sequence: [O J R A] vs. [R A O J] vs. [L E W U] vs. [U L E W]) × 2 (response-to-key order: start with vowel-left and consonant-right vs. start with vowel-right and consonant-left) design, as illustrated in Table 2.

Results

After applying exclusion criteria at the participant level (except for two: minimum of 80% valid "no-go" trials and a minimum of 70% valid "go" trials per cell), 6.44% of the trials were excluded from the analysis (first 5.04% due to incorrect responses, none% because of anticipations with RTs below 250 ms, but another 1.40% within the sequential trimming procedure). The mean RT in the classification task was 615.3 ms (SD = 150.0 ms). Mean RTs per letter and mean accuracies per letter set can be found in the online Supplementary Material B (Tables B1 and B2).

The first regression analysis revealed a significant negative standardised slope for the letter position in the alphabet, $\beta_1 = -0.28$, $t(106) = -3.91$, $p < .001$ (using $\alpha = .0167$). Thus, as expected, and as in Experiments 1 and 2, a significant SAARC effect was found, indicating an increase of right-hand advantage with later positions of the letter in the alphabet (see Figure 4, Panel A). Moreover, the second regression analysis revealed a significant negative standardised slope for the letter position in the learned sequence, $\beta_3 = -0.27$, $t(106) = -3.51$, $p < .001$ (using $\alpha = .0250$). Thus, the expected OPE was found (see Figure 4, Panel B).

Table 3. Data exclusions at the participant level for each experiment.

Exclusion criterion	Exp. 1	Exp. 2	Exp. 3
No serious participation	0	2	1
Other device than desktop, laptop, netbook	0	0	0
Incomplete participation	3	2	3
Noisy environment	2	3	0
Would not use (all) data as experimenter[a]	15	24	16
Fewer than 80% valid 'no-go' trials	0	0	5
Fewer than 70% valid 'go' trials per cell	13	4	9
Incorrect recall of letter sequence at 1st try	-	34	37
Correct recall at 2nd try / 3rd try	-	10 / 7	19 / 5
Incorrect recall even at 3rd try / noted down		17 / 9	13 / 8

[a]Regarding the answers to the question whether participants would use their data if they were the experimenter, across all three experiments, the average accuracy (i.e., proportion of correctly responded "go" trials and correctly non-responded "no-go" trials among all trials) was 0.948 for *yes*, 0.893 for *not all*, and 0.813 *no*, suggesting an accurate self-assessment of by the participants for their own performance relative to others.

Table 4. SAARC and OPE regression slopes resulting from the three experiments, with asterisks indicating significant results (i.e., the respective slope being significantly different from zero using the Bonferroni-Holm corrected significance level).

Experiment	SAARC effect		OPE	
	slope	*p*	slope	*p*
Exp. 1 (control condition)	−0.42	< .001*	-	-
Exp. 2 (maintenance condition)	−0.40	< .001*	−0.08	.142
Exp. 3 (retrieval condition)	−0.28	< .001*	−0.27	< .001*

Discussion

In line with Ginsburg et al. (2014), we found an OPE in this experiment, in which retrieval of the memorised set was required during the classification task (retrieval condition). However, we also found a SAARC effect. Thus, the spatial mapping of letters seems to occur both according to their ordinal sequence in WM (i.e., learned letter sequence) and in LTM (i.e., alphabetical sequence).

Joint exploratory analyses

Comparison of the SAARC effect between letter sets

To examine the SAARC effect in more detail, we also calculated the regressions per letter set (Experiment 1) or learned letter sequence (Experiments 2 and 3) separately. Figure 5 illustrates the resulting average dRTs per letter position in the alphabet and the SAARC slopes for each experiment. All unstandardized and standardised SAARC slopes are descriptively negative. In Experiment 1, the standardised SAARC slopes did not differ significantly between A, J, O, R ($\beta_1 = -0.40$, $n = 75$) and E, L, U, W ($\beta_1 = -0.44$, $n = 74$), when comparing with a two-sided independent-samples *t*-test, $t(147) = 0.27$, $p = .789$. Moreover, in Experiment 2, the standardised SAARC slopes did not differ significantly between the four memorised letter sequences [O J R A] ($\beta_1 = -0.37$, $n = 22$), [R A O J] ($\beta_1 = -0.61$, $n = 22$), [L E W U] ($\beta_1 = -0.36$, $n = 32$), and [U L E W] ($\beta_1 = -0.31$, $n = 29$), when testing the main effect of the between-subject factor letter sequence in a one-way ANOVA, $F(3, 101) = 0.93$, $p = .430$. Similarly, in Experiment 3, the standardised SAARC slopes did not differ significantly between [O J R A] ($\beta_1 = -0.18$, $n = 26$), [R A O J] ($\beta_1 = -0.40$, $n = 25$), [L E W U] ($\beta_1 = -0.49$, $n = 29$), and [U L E W] ($\beta_1 = -0.05$, $n = 27$), when testing the main effect of the between-subject factor letter sequence in a one-way ANOVA, $F(3, 103) = 2.05$, $p = .11$.

Comparison of the OPE between letter sets

To examine the OPE in more detail, we also calculated the regressions per learned letter sequence separately. Figure 6 illustrates the resulting average dRTs per position in the letter sequence and the OPE slopes for Experiments 2 and 3. As reflected by the OPE being significant only in Experiment 3 and not in Experiment 2, the slopes are descriptively more negative in Experiment 3 than in Experiment 2 (cf. Figure 5, Panels B and A, respectively). Interestingly, in contrast to the SAARC slopes, not all OPE slopes are descriptively negative. Moreover, visual inspection suggests that the OPE differs depending on the used letter set (i.e., more negative OPE slopes for E, L, U, W than for A, J, O, R), and more precisely on the learned letter sequence (i.e., positive OPE slopes for [O J R A]). Hence, we conducted a series of additional one-sided *t*-tests with $\alpha = .05$ to test the standardised OPE slopes against zero separately per letter sequence. In Experiment 3, only the OPE for [U L E W] and [L E W U] turned out significant, $t(26) = 5.23$, $p < .001$ and $t(28) = 3.45$, $p = .001$, respectively. None of the tests revealed a significant result in Experiment 2, all $p \geqslant .061$, although the relation between the slopes looks the same. In Experiment 2, the standardised OPE slopes differed significantly between the four memorised letter sequences [O J R A] ($\beta_1 = 0.32$, $n = 22$), [R A O J] ($\beta_1 = -0.10$, $n = 22$), [L E W U] ($\beta_1 = -0.18$, $n = 32$), and [U L E W] ($\beta_1 = -0.27$, $n = 29$), when testing the main effect of the between-subject factor letter sequence in a one-way ANOVA, $F(3, 101) = 2.72$, $p = .048$. Likewise, in Experiment 3, the standardised OPE slopes differed significantly between [O J R A] ($\beta_1 = 0.23$, $n = 26$), [R A O J] ($\beta_1 = -0.09$, $n = 25$), [L E W U] ($\beta_1 = -0.60$,

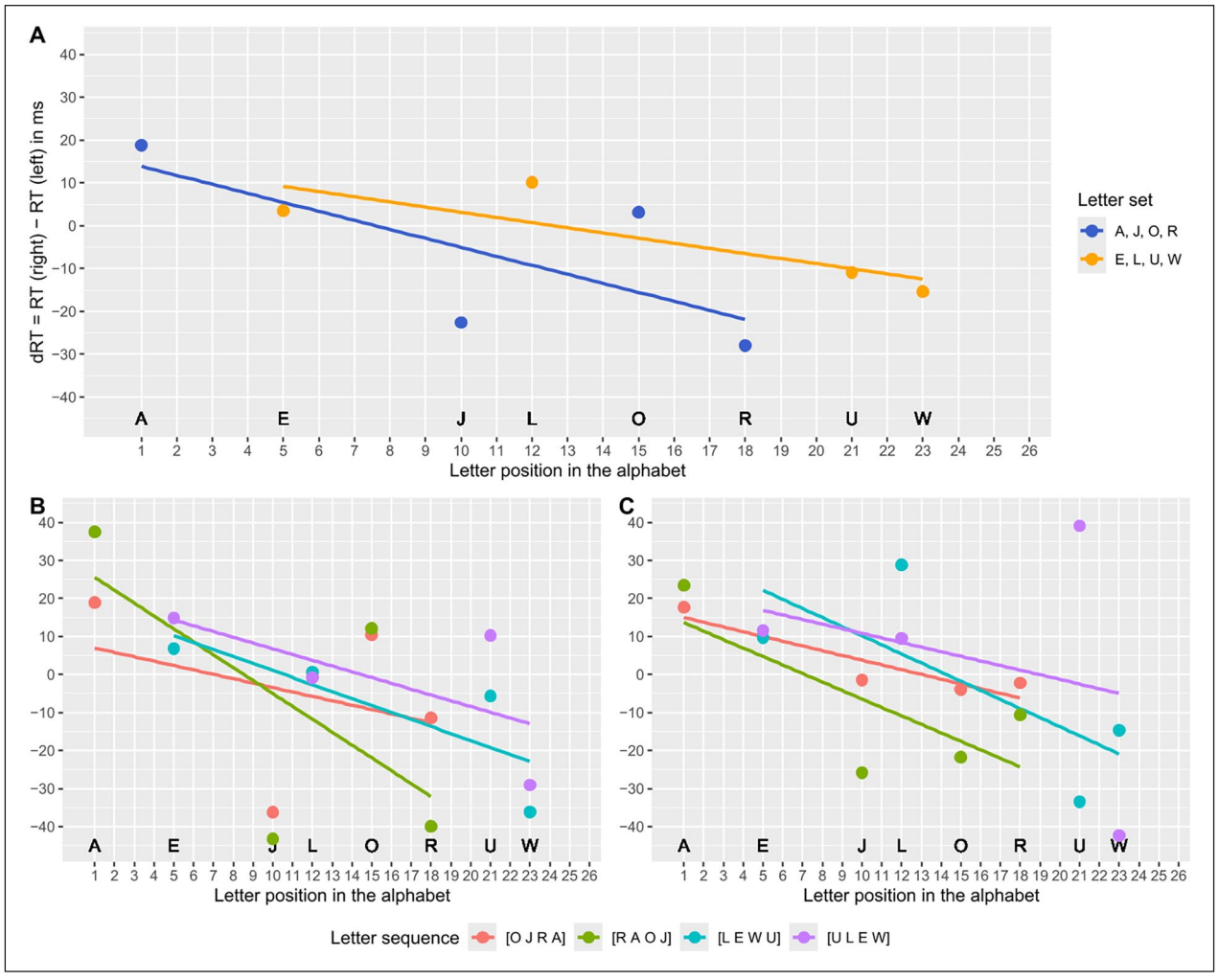


Figure 5. Differences in reaction times (dRT) with the right and left hand depending on the position of the letter in the alphabet (SAARC effect) and averaged over participants in Experiment 1 (control condition, Panel A), Experiment 2 (maintenance condition with intermediate WM salience, Panel B), and Experiment 3 (retrieval condition with high WM salience, Panel C), for each letter set separately.

$n = 29$), and [U L E W] ($\beta_1 = -0.54$, $n = 27$), $F(3, 103) = 8.23$, $p < .001$. Descriptively, the most negative unstandardized slopes were revealed with [U L E W], next [L E W U], then [R A O J], and positive slopes were only revealed with [O J R A] in both Experiments 2 and 3, which is visible in Figure 6. The absence or at least weaker presence of the OPE for [R A O J] and [O J R A] is noticeable. A closer look at Figure 6 can provide an explanation for this: dRTs for the letter A are higher than predicted by the position in the letter sequence for both letter sequences in both experiments and therefore have a high leverage. High dRTs for the letter A indicate that it is judged much faster with the left hand, regardless of its position within the letter sequence, which affects the estimation of the OPE. Note that we decided not to recalculate the OPE and SAARC slopes while excluding the dRT for letter A, because a regression slope should not be based on as few as three datapoints and because letter A only appeared in positions 2 and 4 but not 1 and 3 within the memorised sequences.

Comparison between experiments

In an additional exploratory analysis, we compared the standardised SAARC effect between Experiments 1, 2, and 3, and the standardised OPE between Experiments 2 and 3 (uncorrected $\alpha = .05$). First, a one-sided Jonckheere–Terpstra test for monotone trends was run to find out whether the SAARC effect based on LTM was strongest with no WM salience (Experiment 1), intermediate with intermediate WM salience (Experiment 2), and weakest with highest WM salience (Experiment 3). No evidence was found for this hypothesis, $JT = 22667$, $p = .122$. Second, a one-sided independent-samples Welch t-test was computed to investigate whether the OPE was stronger

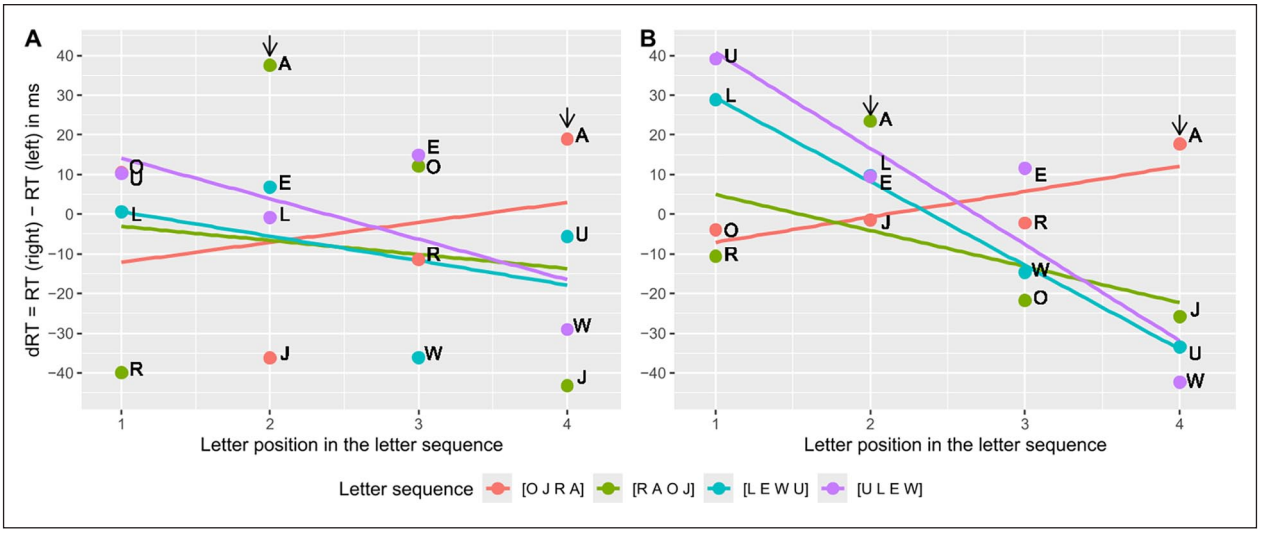


Figure 6. Differences in reaction times (dRT) with the right and left hand depending on the position of the letter in the memorised sequence (OPE) and averaged over participants in Experiment 2 (maintenance condition, Panel A) and Experiment 3 (retrieval condition, Panel B), for each letter set separately. The outlier letter A is marked with an arrow.

with high WM salience (Experiment 3) than with intermediate WM salience (Experiment 2). Our data supported this hypothesis, $t(209.65) = 1.67$, $p = .049$. In addition, in one ANOVA per letter sequence, we explored whether WM salience affected spatial response preferences for different letters to different extents, but the interaction for dRTs between experiments and letters was not significant in any of the four-letter sequences (see the online Supplementary Material D).

Correlation between the SAARC effect and the OPE

Moreover, we investigated whether the SAARC effect and the OPE correlate with one another. Our data revealed no evidence of a correlation between the effects. For Experiment 2, the correlation remained non-significant, $r = 0.02$, $t(103) = 0.21$, $p = .832$, and so did the correlation for Experiment 3, $r = 0.07$, $t(105) = 0.77$, $p = .443$.

General discussion

With this series of experiments, we investigated whether the account postulating a stable spatial long-term representation and the WM account postulating spatial short-term representations (van Dijck et al., 2014) can be transferred from numbers to other ordered stimuli such as letters. Previous literature has shown evidence for both accounts, and the SNARC effect and the OPE for numbers have even been observed simultaneously. Assuming similar mechanisms for mental spatial representations of letters as for those of numbers, a SAARC reflecting the alphabetical order stored in LTM was expected when no new ordinal sequence had to be learned or when a new ordinal sequence merely had to be maintained without being task-relevant. In contrast, an OPE for a new ordinal sequence stored in WM was hypothesised only when the sequence needed to be retrieved from WM to perform the task correctly.

The starting point was a control condition (Experiment 1), with which we wanted to test the suitability of our go/no-go design for the investigation of spatial response preferences for letters. Importantly, as compared to numbers which comprise information on both magnitude and order, letters have an ordered nature but no inherent magnitude characteristic. In line with previous results of Gevers et al. (2003), we found that participants responded faster to letters at the beginning of the alphabet with the left hand than with the right hand, while letters at the end of the alphabet were responded to faster with the right than with the left hand (i.e., a SAARC effect), despite the presence of "no-go" trials. This manipulation check provided a solid basis for the investigation of a maintenance condition (i.e., intermediate WM salience, Experiment 2) vs. a retrieval condition (i.e., high WM salience, Experiment 3) of previously memorised sequences during consonant-vowel classification on spatial associations. Here, on top of the inherent ordinality letters stemming from the alphabet stored in LTM, an artificial different ordinality was created in WM by the instruction to memorise a letter sequence. In line with the results of Ginsburg et al. (2014), in Experiment 2, we found no spatial response preference for the position of the letters in the memorised sequence (i.e., no evidence for an OPE). However, as in Experiment 1, we found a response preference for the position of the letters within the alphabet (i.e., a SAARC effect). In this case, the findings for numbers seem to be transferable to letters. Furthermore, also consistent with results by Ginsburg et al. (2014), we found an OPE in Experiment 3, in which

the WM salience was high. Moreover, our exploratory analysis confirmed that the OPE was significantly stronger with high than with intermediate WM salience. This supports the conclusion of Ginsburg et al. (2014) and van Dijck et al. (2014) that retrieval during the task is crucial for the presence of an OPE and extends it to classification tasks of other ordinal stimuli than numbers, namely letters. Interestingly, however, one needs to consider that the ordinal position stored in LTM and thus the spatial mapping of some letters might be very robust and not be overruled by newly learned ordinal positions stored in WM.

These results have important implications for theories and explanations of spatial-numerical associations on the one hand and of spatial associations with non-numerical ordinal stimuli on the other. First, they provide evidence that ordinality (short- and long-term) can elicit spatial mapping in case no cardinality information is present. When participants need to memorise a sequence of arbitrary objects in WM (e.g., fruit and vegetable names; van Dijck & Fias, 2011) and recall the sequence during task execution, the WM ordering is decisive (i.e., there is no magnitude nor long-term memory order). In the case of overlearned ordinal sequences such as letters, the ordinality is crucial (i.e., there is no quantity or magnitude). Whether a sequence kept in LTM or a sequence kept in WM is decisive for the spatial mapping (i.e., when both long- and short-term ordinality are present) remained unclear. In the present study, we found that the salience of the WM order is decisive.

The SAARC effect reflecting spatial mappings of ordinality coming from LTM seems to be small. Specifically, the SAARC effect found by Gevers et al. (2003) had an unstandardized slope of 1.12 ms. The unstandardized slopes in the current study were similar (i.e., −1.65 ms in Experiment 1, −1.93 ms in Experiment 2, and −1.78 ms in Experiment 3). This might even have led Dehaene and colleagues (1993) to rule out the role of ordinality in LTM for creating spatial-numerical associations (see their Experiment 4). Due to its rather small effect sizes, a true SAARC effect might have simply gone undetected with only 10 participants. However, the spatial response preferences we found for letters in Experiment 1 support the hypothesis that spatial associations could be the result of overlearned sequences and thus not be unique to numerical stimuli. Our results are in line with the results by Gevers et al. (2003) showing spatial response preferences for non-numerical ordinal stimuli, namely for letters or months.

At first sight, the present results support both the LTM and the WM account (van Dijck & Fias, 2011; see also Abrahamse et al., 2014, 2016; van Dijck et al., 2014). The WM account assumes that spatial associations are the result of the ordinal position of the stimuli in the WM, with stimuli at the beginning of the memorised sequence being associated with the left and stimuli at the end of the sequence being associated with the right. In the present study, an OPE was detected (i.e., a spatial response preference depending on the ordinal position of the stimuli in WM), investigated via letter sequences memorised before the consonant-vowel classification. This was analogous to van Dijck and Fias' original study design, with number sequences memorised before parity judgement, but opposed to Koch et al.'s (2023) design, where inherent ordinal information of numbers was used, and no number sequences had to be memorised before the task. Since not only numerical stimuli can be memorised in an artificial order in WM, the WM approach can explain not only the SNARC effect but also spatial associations of non-numerical ordinal stimuli, like in our case letters. However, we only found the OPE if the memorised letters had to be retrieved during the consonant-vowel classification task (retrieval condition) but not if they were only held in memory for later recall (maintenance condition), although we found a SAARC effect reflecting long-term associations. This pattern of results is consistent with the results of Ginsburg et al. (2014) who identified retrieval during the classification task as a prerequisite for the occurrence of the OPE and concluded that the spatial associations between the ordinal stimuli in WM and response arise during the classification itself when the information is retrieved. Importantly, these findings can be interpreted in two ways: First, when no retrieval of the WM content is required by the task, this could either result in the ordinality information of the WM content not being activated, while the stored items are activated. Second, it could result in the WM content not being activated at all and being only passively stored.

Interestingly, we also found a coexistence of the OPE and the SAARC effect in the retrieval condition (Experiment 3). These results are in line with studies by Ginsburg and Gevers (2015) and Huber et al. (2016), who have already shown the coexistence of the OPE and the SNARC effect for numerical stimuli. The two simultaneously arising effects have been interpreted as the result of different representations, with the SNARC effect being based on the representation of numbers and their respective canonical order in LTM, and the OPE being based on the spatial short-term bindings in WM, which are formed during retrieval (Ginsburg and Gevers, 2015; Huber et al., 2016; van Dijck et al., 2015). This hybrid explanatory approach for numbers seems to be applicable to non-numerical ordered stimuli such as letters and can therefore also account for our results: The SAARC effect, which is the result of the alphabetical order stored in LTM, can be overcome by an OPE, which reflects short-term spatial associations based on sequences stored in WM if certain conditions are fulfilled (e.g., necessity of retrieval of the learned sequence during the task). Note that the coexistence of the two effects within one experiment could be driven by different trials instead of them influencing response patterns simultaneously. As recently demonstrated by Ftaïta et al. (2024), spatialization of ordinal

information coming from a newly learned sequence stored in WM vanishes over time when the sequence is repeated. A potential explanation offered by the authors is that spatialization gets weaker and finally disappears with the newly learned sequence being chunked and moved from WM to LTM. In our design, where the encoding phase is followed by a relatively long classification phase, the OPE could be dominant in the first trials, where the newly learned sequence is still very present in WM, and the SAARC could be dominant in the later trials, where the sequence fades into the background. This would be in line with Ftaïta et al.'s findings; however, it is not possible to investigate this in the present study. In both their and our study, the response-to-key assignment was switched after the first half, and the OPE was calculated as a regression of the differences in RTs between left and right responses on the ordinal position. Crucially, Ftaïta et al. introduced a new ordinal sequence before the second half of their experiment, which is why both left and right responses might equally be affected by spatialization based on ordinal positions. In contrast, in the present study, only one sequence had to be remembered for both response-to-key assignments together, which is why the spatialization might only have occurred in the first trials with one response key but not in the first trials with the respective other. Although for these reasons changes in the OPE over the course of the experiment cannot be tested within the present study, another argument for the OPE vanishing over time in Ftaïta et al.'s study is that it was only task-relevant *which* elements were part of the memorised sequence, whereas their order was not task-relevant. This is an important difference to van Opstal et al.'s (2009) experiment and to two of Previtali et al.'s (2010) experiments, where participants were to make order judgements. Hence, the ordinal information might get lost over time, which is not the case for the canonical order stored in LTM.

Stimulus-specific effects on short-term and long-term letter-space associations

Regarding numbers, it has been known for a long time that zero has a special role for spatial-numerical associations (Nuerk et al., 2004; Pinhas & Tzelgov, 2012). Its mental representation differs from those of other numbers (Nieder, 2016), leading to observations such as its parity status being unclear to many students (e.g., Levenson et al., 2007). Uncertainty in parity judgement for number zero can therefore impair the estimation of the SNARC effect (Huber et al., 2016; Nuerk et al., 2004; Roth et al., 2025b). Our results suggest that such stimulus-specific effects can be found for letters as well and that they need to be considered in interpretations. Some letters seem to have a very strong long-term spatial association so that short-term associations do not play a role for them. At the same time, for many other letters, associations in both LTM and WM seem to influence responses. As the WM account did not propose different statuses of different elements within a specific sequence, our design was not prepared to test such effects in a systematic fashion. However, our observations call for such investigations in future studies.

From what we could see in our study, responses were descriptively faster for the letter A than for any other letter (see Table B1 in the online Supplementary Material B). Moreover, the letter A was answered faster with the left hand than with the right hand regardless of its position in the memorised sequence. We can speculate that this pattern could be explained by an especially strong, immediately available long-term representation of the letter A as the first letter of the alphabet in LTM. This leading to the respective spatial association and response preference with the left side might have exceeded the interfering activation of the spatial association and response preference based on the short-term representation of the memorised letter sequence in WM. Faster responses with the left for the letter A, regardless of its position in the memorised letter sequence, might be the reason for the OPE to differ between letter sequences, as can be seen in Figure 6 in the reversed slope for [O J R A]. Note that Gevers et al. (2003) did not use letter A in their study. Our findings could indicate that the OPE can be overridden by automated, highly overlearned stimulus associations. As outlined above, the influence of individual stimuli on spatial response behaviour has also been observed for numbers before. These data are the first to show such stimulus-specific influences for spatial LTM and WM associations for letters. Future research should investigate to what extent overlearned associations from items stored in the LTM can override short-term associations relying on sequences stored in WM on the one hand, and to what extent short-term associations can override long-term associations on the other hand. For instance, are influences of spatial long-term associations and seemingly facilitated processing limited to A as the first letter of the alphabet, or is there an analogous effect for the letter Z as the last letter of the alphabet? Here, one would expect a strong spatial association with the right. It would also be interesting to see whether a letter-specific effect is also evident for neighbouring letters such as B and C or X and Y. Moreover, training studies could investigate how newly created but strongly trained associations affect the OPE.

The fact that some letters seem to have rather stable LTM representations, whereas LTM representations can be more easily overruled by WM content for other letters, can also explain that no correlation was found between the OPE and the SAARC effect. However, apart from stimulus-specific effects, there are at least two more reasons for the lack of evidence for a correlation: The experimental task might have low test–retest reliability, or the OPE and the SAARC effect might just vary and be unstable on an intraindividual

level. This seems to be the case for the SNARC effect (Roth et al., 2024) and be similar for the OPE and SAARC effect.

Modal and amodal representations of numbers, letters, and their spatialization

Kaup et al. (2024) suggest a major distinction of mental representations between modal (i.e., modality-specific) and amodal (i.e., modality-unspecific) representations. They also suggest that explanatory approaches for the SNARC effect can be categorised into modal and amodal approaches. According to the view proposed by Dehaene et al. (1993), the SNARC effect is seen as the result of modal magnitude representations specific to numbers, which are arranged on the MNL aligned from left-to-right in Western cultures due to modal experiences of reading and writing direction. According to ATOM (Bueti & Walsh, 2009; Walsh, 2003), this modal account magnitude dimensions other than numbers (e.g., physical size or luminance; Cohen Kadosh et al., 2008; Fumarola et al., 2014; Ren et al., 2011). According to Pitt and Casasanto (2018), this holds only for the modality of prothetic dimensions (i.e., where humans experience quantitative variations) that entail quantity or magnitude, but not for metathetic dimensions (i.e., where humans experience qualitative variations). Conversely, the WM account by van Dijck and Fias (2011) can be seen as an amodal (i.e., unspecific and not only applicable to quantity or magnitude dimensions, but also other items) approach, as it can explain spatial response preferences for very different kinds of stimuli, including objects without any inherent magnitude or ordinality structure, such as fruits and vegetables.

The current data, especially from Experiment 3, confirm the notion that both modal and amodal representations can contribute to letter-space associations, similar to spatial-numerical associations (see also Koch et al., 2023, for numbers). Notably, the current data suggest that the influence of modal and amodal representations on the spatial mental mapping of letters may also depend on individual letter identity and learning. For example, a strong association of A with the left could occur because this letter being the first in the alphabet is highly overlearned, both visually and auditorily. For instance, in German, learning the alphabet at school is colloquially usually referred to as "learning the ABC," so A comes first even in the colloquial name of the alphabet. This highly overlearned standard visual and auditory input and output may explain the strong long-term associations and the weak or even non-existent short-term association for the letter A. Likewise, a recent study has shown that the spatial association of number 1 with the left is stronger than what would be expected when deriving predictions from the regression line fit to all numbers within the stimulus set where number 1 is the smallest number, or when comparing with the spatial association of the lowest number in other stimulus sets (Roth et al., 2025b). Future studies should pit spatial associations resulting from WM or LTM against each other by administering sequences to be memorised by participants in which the stimuli have the opposite artificial order as compared to their canonical order.

Limitations of the present study

Despite our recruitment efforts, the sample size required according to the preregistered power analysis was not achieved. However, 169 participants took part in the maintenance condition (Experiment 2) and another 168 in the retrieval condition (Experiment 3), which largely exceeds the sample sizes used within the published literature testing the OPE (e.g., Ginsburg et al., 2014; Ginsburg & Gevers, 2015; van Dijck & Fias, 2011) and the SAARC effect (Gevers et al., 2003). This still makes our samples the largest that has ever been used to investigate the OPE with letters and the SAARC effect (note, however, that previous studies have administered several ordinal sequences to be memorised, whereas there was only one such sequence in our experiments). These sample sizes permitted us to observe a significant OPE at least in Experiment 3 and a significant SAARC effect in all three Experiments (see Table 4).

The present study was conducted online, which means that—in contrast to a laboratory study—it was not possible to keep all conditions constant for all participants. To be able to exclude data from participation with distractions (such as noise), quality items were assessed at the end of the study. Based on these quality items, we excluded datasets from participants who indicated that they performed the experiment in a noisy environment or that they would not use their data if they were the experimenters. Importantly, the SNARC effect can be observed reliably in online settings and is comparable to laboratory settings (e.g., Cipora et al., 2019; Gökaydin et al., 2018; Koch et al., 2023; Roth et al., 2025b; Roth et al., 2025a), so we assume that this is the case for the OPE and the SAARC effect, too.

We also asked the participants how they memorised the letter sequence and whether they had noted it down, implying that they had not necessarily stored the letter sequence in their WM but rather externalised it. In case participants noted down the letter sequence, neither maintenance of the memorised letter sequence nor retrieval from WM during the classification task can be ensured, which is why respective participants were excluded. In addition to these potential difficulties, the online setting does offer various advantages, such as the collection of a larger and more diverse sample. This increases representativeness and generalisability of the results as compared to standard laboratory experiments with university students (Reips, 2002). Nevertheless, it should be mentioned that the present results are limited to cultures with a left-to-right reading and writing direction, which cannot easily be transferred to cultures with a different reading and

writing directions (Bulut et al., 2025; Hochman et al., 2025; Shaki et al., 2009; Shaki & Gevers, 2011; Zebian, 2005).

Importantly and as opposed to many previous studies, there was only one encoding phase in Experiments 2 and 3 of the present study, such that participants completed 320 trials with the same letters determining the go/no-go rule. One could argue that after a certain number of trials in Experiment 3, participants might have generated an episodic memory trace of the task execution and know which letters to respond to and which not. That is, they might not need to retrieve the ordered sequence stored in WM anymore after a certain number of trials so that the order of the sequence does not affect the response patterns anymore in later trials. Moreover, the position of each letter in LTM and in WM was not fully orthogonal in the present study, because all letters have a fixed position in LTM but were only presented on two instead of four potential positions within the ordinal sequence stored in WM (e.g., the letter O was on position 1 in participants administered with [O J R A] and on position 3 for participants administered with [R A O J], but never on positions 2 and 4). Therefore, the current design is not ideal for the investigation of the interactions between letter positions in LTM and WM. However, 20 repetitions per combination of stimulus and response-to-key assignment permitted us a very precise estimate of the SAARC effect and of the OPE, which was a clear advantage of the current study's design with a relatively long classification phase.

Conclusion

The current study demonstrates the coexistence of long-term letter-space associations (i.e., stimulus-specific for the letter A and the SAARC effect) and short-term stimulus-unspecific associations (i.e., the OPE produced by a letter sequence stored in WM). These two representations do not seem to be mutually exclusive (see also Ginsburg & Gevers, 2015). While the first point towards modal spatial mappings for letters (i.e., letters early in the alphabet are associated with the left and letters late in the alphabet are associated with the right), the latter point towards amodal stimulus-unspecific spatial mappings (i.e., stimuli early in a memorised sequence are associated with the left and stimuli late in a memorised sequence are associated with the right). Importantly, not only did we thereby replicate the findings by Gevers et al. (2003), but we were also able to extend the findings by Ginsburg et al. (2014) from numbers to letters as a different type of ordinal stimuli. We found that mere maintenance of a memorised sequence is not sufficient to observe an OPE; instead, retrieval during the classification task is crucial. Our findings point towards the possibility that overlearned orders and their spatial associations in LTM can coexist with newly learned orders and their spatial short-term associations in WM, thus providing evidence for a hybrid model that integrates both modal and amodal representations.

Acknowledgements

The authors wish to thank Elena Albu for giving feedback on the manuscript prior to submission, Jonas Honecker for his help with the data collection, and Sebastian Sandbrink for English language proofreading of the manuscript.

Declaration of conflicting interests

The authors declared no conflicts of interest with respect to the research, authorship, or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the Deutsche Forschungsgemeinschaft (DFG—German Research Foundation) within the Research Unit FOR2718: "Modal and Amodal Cognition" (project numbers BU 1335/12-1 granted to MVB and NU 265/5-1 and NU 265/5-2 granted to HCN). JFH was supported by the DFG project NU265/5-1 and LR by the DFG project NU 265/8-1. KC was supported by the UKRI Economic and Social Research Council (grant number ES/W002914/1).

ORCID IDs

Lilly Roth https://orcid.org/0000-0002-2690-0504
Jean-Philippe van Dijck https://orcid.org/0000-0003-0378-3891
Krzysztof Cipora https://orcid.org/0000-0003-0077-9336
Martin V. Butz https://orcid.org/0000-0002-8120-8537
Hans-Christoph Nuerk https://orcid.org/0000-0002-0331-7498

Data accessibility statement

The data from the present experiment are publicly available at the Open Science Framework website: https://osf.io/v84m5/.

Supplementary material

The Supplementary Material is available at qjep.sagepub.com

Notes

1. A recent study by Vivion et al. (2024) showed that sequences of four items are most suitable to detect spatialization.
2. Bonferroni correction was preregistered, but we decided to apply Bonferroni-Holm correction, because it represents a better balance between type I and type II errors and is a generally more recommended method. At the same time, in the present study, the significance of all results would remain unchanged if the Bonferroni method was used.
3. When excluding this participant from the main analyses, the results (see below) did not change: The SAARC effect remained significant, $\beta_1 = 0.38$, $t(103) = 5.95$, $p < .001$, and

the OPE remained non-significant, $\beta_3 = 0.09$, $t(103) = 1.11$, $p = .136$. Hence, we did not exclude this participant's data from any data analyses reported in this manuscript.

4. When excluding this participant from the main analyses, the results (see below) did not change: The SAARC effect, $\beta_1 = 0.28$, $t(105) = 3.87$, $p < .001$, and the OPE remained significant, $\beta_3 = 0.27$, $t(105) = 3.55$, $p < .001$. Hence, we did not exclude this participant's data from any data analyses reported in this manuscript.

References

Abrahamse, E. L., & van Dijck, J.-P. (2023). Ranking-space: Magnitude makes sense through spatially scaffolded ranking. *Frontiers in Psychology*, *14*, Article 1224254. https://doi.org/10.3389/fpsyg.2023.1224254

Abrahamse, E. L., van Dijck, J.-P., & Fias, W. (2016). How does working memory enable number-induced spatial biases? *Frontiers in Psychology*, *7*, Article 977. https://doi.org/10.3389/fpsyg.2016.00977

Abrahamse, E. L., van Dijck, J.-P., Majerus, S., & Fias, W. (2014). Finding the answer in space: The mental whiteboard hypothesis on serial order in working memory. *Frontiers in Human Neuroscience*, *8*, 1–12. https://doi.org/10.3389/fnhum.2014.00932

Bächtold, D., Baumüller, M., & Brugger, P. (1998). Stimulus-response compatibility in representational space. *Neuropsychologia*, *36*(8), 731–735. https://doi.org/10.1016/S0028-3932(98)00002-5

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), Article 16. https://doi.org/10.5334/joc.72

Bueti, D., & Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical Transactions of the Royal Society B, Biological Sciences*, *364*(1525), 1831–1840. https://doi.org/10.1098/rstb.2009.0028

Bulut, M., Roth, L., Bahreini, N., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2025). One direction? Cultural aspects of the mental number line beyond reading direction. *Psychological Research*, *89*, Article 37. https://doi.org/10.1007/s00426-024-02038-4

Casasanto, D., & Pitt, B. (2019). The faulty magnitude detector: Why SNARC-like tasks cannot support a Generalized Magnitude System. *Cognitive Science*, *43*(10), Article e12794. https://doi.org/10.1111/cogs.12794

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & Rosario, H. D. (2020). *pwr: Basic functions for power analysis* [R package, Version 1.3-0]. https://CRAN.R-project.org/package=pwr

Cipora, K., Haman, M., Domahs, F., & Nuerk, H.-C. (2020). Editorial: On the development of space-number relations: Linguistic and cognitive determinants, influences, and associations. *Frontiers in Psychology*, *11*, Article 182. https://doi.org/10.3389/fpsyg.2020.00182

Cipora, K., Patro, K., & Nuerk, H.-C. (2018). Situated Influences on Spatial–Numerical Associations. In T. L. Hubbard (Ed.), *Spatial biases in perception and cognition* (pp. 41–59). Cambridge University Press. https://doi.org/10.1017/9781316651247.004

Cipora, K., Soltanlou, M., Reips, U. D., & Nuerk, H. C. (2019). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behavior Research Methods*, *51*, 1676–1692. https://doi.org/10.3758/s13428-019-01213-5

Cohen Kadosh, R., Cohen Kadosh, K., & Henik, A. (2008). When brightness counts: The neuronal correlate of numerical–luminance interference. *Cerebral Cortex*, *18*(2), 337–343. https://doi.org/10.1093/cercor/bhm058

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371–396. https://doi.org/10.1037/0096-3445.122.3.371

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 626–641. https://doi.org/10.1037/0096-1523.16.3.626

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Fias, W., Brysbaert, M., Geypens, F., & d'Ydewalle, G. (1996). The importance of magnitude information in numerical processing: Evidence from the SNARC effect. *Mathematical Cognition*, *2*(1), 95–110. https://doi.org/10.1080/135467996387552

Fias, W., Lauwereyns, J., & Lammertyn, J. (2001). Irrelevant digits affect feature-based attention depending on the overlap of neural circuits. *Cognitive Brain Research*, *12*(3), 415–423. https://doi.org/10.1016/S0926-6410(01)00078-7

Fischer, M. H., Mills, R. A., & Shaki, S. (2010). How to cook a SNARC: Number placement in text rapidly changes spatial-numerical associations. *Brain and Cognition*, *72*(3), 333–336. https://doi.org/10.1016/j.bandc.2009.10.010

Ftaïta, M., Guida, A., Fartoukh, M., & Mathy, F. (2024). Spatial–positional associations in short-term memory can vanish in long-term memory. *Memory & Cognition*. https://doi.org/10.3758/s13421-024-01577-w

Fumarola, A., Prpic, V., Da Pos, O., Murgia, M., Umiltà, C., & Agostini, T. (2014). Automatic spatial association for luminance. *Attention, Perception, & Psychophysics*, *76*(3), 759–765. https://doi.org/10.3758/s13414-013-0614-y

Gevers, W., Reynvoet, B., & Fias, W. (2003). The mental representation of ordinal sequences is spatially organized. *Cognition*, *87*(3), 87–95. https://doi.org/10.1016/S0010-0277(02)00234-2

Gevers, W., Reynvoet, B., & Fias, W. (2004). The mental representation of ordinal sequences is spatially organized: Evidence from days of the week. *Cortex*, *40*(1), 171–172. https://doi.org/10.1016/S0010-9452(08)70938-9

Giner-Sorolla, R., Montoya, A. K., Reifman, A., Carpenter, T., Lewis, N. A., Aberson, C. L., Bostyn, D. H., Conrique, B. G., Ng, B. W., Schoemann, A. M., & Soderberg, C. (2024). Power to detect what? Considerations for planning and evaluating sample size. *Personality and Social Psychology Review*, *28*(3), 276–301. https://doi.org/10.1177/10888683241228328

Ginsburg, V., & Gevers, W. (2015). Spatial coding of ordinal information in short-and long-term memory. *Frontiers in Human Neuroscience*, *9*, Article 8. https://doi.org/10.3389/fnhum.2015.00008

Ginsburg, V., van Dijck, J.-P., Previtali, P., Fias, W., & Gevers, W. (2014). The impact of verbal working memory on number-space associations. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(4), 976–986. https://doi.org/10.1037/a0036378

Gökaydin, D., Brugger, P., & Loetscher, T. (2018). Sequential effects in SNARC. *Scientific Reports*, *8*(1), Article 10996. https://doi.org/10.1038/s41598-018-29337-2

Guida, A., & Campitelli, G. (2019). Explaining the SPoARC and SNARC effects with knowledge structures: An expertise account. *Psychonomic Bulletin & Review*, *26*(2), 434–451. https://doi.org/10.3758/s13423-019-01582-0

Guida, A., & Lavielle-Guida, M. (2014). 2011 space odyssey: Spatialization as a mechanism to code order allows a close encounter between memory expertise and classic immediate memory studies. *Frontiers in Psychology*, *5*, Article 573. https://doi.org/10.3389/fpsyg.2014.00573

Herrera, A., Macizo, P., & Semenza, C. (2008). The role of working memory in the association between number magnitude and space. *Acta Psychologica*, *128*(2), 225–237. https://doi.org/10.1016/j.actpsy.2008.01.002

Hochman, S., Havedanloo, R., Heysieattalab, S., & Soltanlou, M. (2025). How does language modulate the association between number and space? A registered report of a cross-cultural study of the spatial–numerical association of response codes effect. *Journal of Experimental Psychology: General*, *154*(2), 305–324. https://doi.org/10.1037/xge0001653

Huber, S., Klein, E., Moeller, K., & Willmes, K. (2016). Spatial-numerical and ordinal positional associations coexist in parallel. *Frontiers in Psychology*, 7, Article 438. https://doi.org/10.3389/fpsyg.2016.00438

Kassambara, A. (2023). *ggpubr: "ggplot2" based publication ready plots* [R package, version 0.6.0]. https://CRAN.R-project.org/package=ggpubr

Kaup, B., Ulrich, R., Bausenhart, K. M., Bryce, D., Butz, M. V., Dignath, D., Dudschig, C., Franz, V. H., Friedrich, C., Gawrilow, C., Heller, J., Huff, M., Hütter, M., Janczyk, M., Leuthold, H., Mallot, H., Nuerk, H.-C., Ramscar, M., Said, N., Svaldi, S., Wong, H. Y. (2024). Modal and amodal cognition: An overarching principle in various domains of psychology. *Psychological Research*, *88*(2), 307–337. https://doi.org/10.1007/s00426-023-01878-w

Koch, N. N., Huber, J. F., Lohmann, J., Cipora, K., Butz, M. V., & Nuerk, H.-C. (2023). Mental number representations are spatially mapped both by their magnitudes and ordinal positions. *Collabra: Psychology*, *9*(1), Article 67908. https://doi.org/10.1525/collabra.67908

Leiner, D. J. (2021). *SoSci survey* [Computer software, Version 3.2.23). https://www.soscisurvey.de

Levenson, E., Tsamir, P., & Tirosh, D. (2007). Neither even nor odd: Sixth grade students' dilemmas regarding the parity of zero. *The Journal of Mathematical Behavior*, *26*(2), 83–95. https://doi.org/10.1016/j.jmathb.2007.05.004

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 149–157. https://doi.org/10.1037/0278-7393.16.1.149

Mingolo, S., Prpic, V., Mariconda, A., Brugger, P., Drack, T., Bilotta, E., Agostini, T., & Murgia, M. (2024). It's SNARC o' clock: Manipulating the salience of the context in a conceptual replication of Bächtold et al.'s (1998) clockface study. *Psychological Research*, *88*, 837–851. https://doi.org/10.1007/s00426-023-01893-x

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–1520. https://doi.org/10.1038/2151519a0

Nieder, A. (2016). Representing something out of nothing: The dawning of zero. *Trends in Cognitive Sciences*, *20*(11), 830–842. https://doi.org/10.1016/j.tics.2016.08.008

Nuerk, H. C., Iversen, W., & Willmes, K. (2004). Notational modulation of the SNARC and the MARC (linguistic markedness of response codes) effect. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *57*(5), 835–863. https://doi.org/10.1080/02724980343000512

Nuerk, H. C., Wood, G., & Willmes, K. (2005). The universal SNARC effect: The association between number magnitude and space is amodal. *Experimental Psychology*, *52*(3), 187–194. https://doi.org/10.1027/1618-3169.52.3.187

Pavlovia Surveys (Open Science Tools). (2021). https://pavlovia.org/

Pinhas, M., & Tzelgov, J. (2012). Expanding on the mental number line: Zero is perceived as the "smallest." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1187–1205. https://doi.org/10.1037/a0027390

Pitt, B., & Casasanto, D. (2018). Spatializing emotion: No evidence for a domain-general magnitude system. *Cognitive Science*, *42*(7), 2150–2180. https://doi.org/10.1111/cogs.12568

Previtali, P., De Hevia, M. D., & Girelli, L. (2010). Placing order in space: The SNARC effect in serial learning. *Experimental Brain Research*, *201*(3), 599–605. https://doi.org/10.1007/s00221-009-2063-3

Prpic, V., Fumarola, A., Tommaso, M., de Luccio, R., Murgia, M., & Agostini, T. (2016). Separate mechanisms for magnitude and order processing in the spatial-numerical association of response codes (SNARC) effect: The strange case of musical note values. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1241–1251. https://doi.org/10.1037/xhp0000217

Prpic, V., Mingolo, S., Agostini, T., & Murgia, M. (2021). Magnitude and order are both relevant in SNARC and SNARC-like effects: A commentary on Casasanto and Pitt (2019). *Cognitive Science*, *45*(7), Article e13006. https://doi.org/10.1111/cogs.13006

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software, Version 4.3.2]. R Foundation for Statistical Computing. https://www.R-project.org/

Reips, U. D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. https://doi.org/10.3969/j.issn.1673-8225.2010.12.045

Ren, P., Nicholls, M. E. R., Ma, Y., & Chen, L. (2011). Size matters: Non-numerical magnitude affects the spatial coding

of response. *PLOS ONE*, *6*(8), Article e23553. https://doi.org/10.1371/journal.pone.0023553

Restle, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology*, *83*(2, Pt. 1), 274–278. https://doi.org/10.1037/h0028573

Roth, L., Caffier, J., Cipora, K., Reips, U.-D., & Nuerk, H.-C. (2025a). True colors SNARC: Semantic number processing is highly automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001431

Roth, L., Caffier, J., Reips, U.-D., Nuerk, H.-C., Overlander, A. T., & Cipora, K. (2025b). One and only SNARC? Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude. *Royal Society Open Science*, *12*(1), Article 241585. https://doi.org/10.1098/rsos.241585

Roth, L., Jordan, V., Schwarz, S., Willmes, K., Nuerk, H.-C., van Dijck, J.-P., & Cipora, K. (2024). Don't SNARC me now! Intraindividual variability of cognitive phenomena—Insights from the Ironman paradigm. *Cognition*, *248*, Article 105781. https://doi.org/10.1016/j.cognition.2024.105781

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017a). Space in numerical and ordinal information: A common construct? *Journal of Numerical Cognition*, *3*(2), 164–181. https://doi.org/10.5964/jnc.v3i2.40

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017b). Switching between multiple codes of SNARC-like associations: Two conceptual replication attempts with anodal tDCS in sham-controlled cross-over design. *Frontiers in Neuroscience*, *11*, Article 654. https://doi.org/10.3389/fnins.2017.00654

Seshan, V. E., & Whiting, K. (2023). *clinfun: Clinical trial design and data analysis functions* [R package, Version 1.1.5]. https://CRAN.R-project.org/package=clinfun

Shaki, S., Fischer, M. H., & Petrusic, W. M. (2009). Reading habits for both words and numbers contribute to the SNARC effect. *Psychonomic Bulletin and Review*, *16*(2), 328–331. https://doi.org/10.3758/PBR.16.2.328

Shaki, S., & Gevers, W. (2011). Cultural characteristics dissociate magnitude and ordinal information processing. *Journal of Cross-Cultural Psychology*, *42*(4), 639–650. https://doi.org/10.1177/0022022111406100

van Dijck, J.-P., Abrahamse, E. L., Acar, F., Ketels, B., & Fias, W. (2014). A working memory account of the interaction between numbers and spatial attention. *Quarterly Journal of Experimental Psychology*, *67*(8), 1500–1513. https://doi.org/10.1080/17470218.2014.903984

The Jamovi Project. (2023). *jamovi* [Computer software, Version 2.3]. https://www.jamovi.org

van Dijck, J.-P., & Fias, W. (2011). A working memory account for spatial–numerical associations. *Cognition*, *119*(1), 114–119. https://doi.org/10.1016/j.cognition.2010.12.013

van Dijck, J.-P., Gevers, W., & Fias, W. (2009). Numbers are associated with different types of spatial information depending on the task. *Cognition*, *113*(2), 248–253. https://doi.org/10.1016/j.cognition.2009.08.005

van Dijck, J.-P., Ginsburg, V., Girelli, L., & Gevers, W. (2015). Linking numbers to space: From the mental number line towards a hybrid account. In R. C. Kadosch & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 89–105). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.020

Van Opstal, F., Fias, W., Peigneux, P., & Verguts, T. (2009). The neural representation of extensively trained ordered sequences. *NeuroImage*, *47*(1), 367–375. https://doi.org/10.1016/j.neuroimage.2009.04.035

Vivion, M., Guida, A., Ramanoel, S., & Mathy, F. (2024). *Spatial-positional association of response codes is modulated by the number of items in working memory* [preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/3z2gs

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. https://doi.org/10.1016/j.tics.2003.09.002

Wickham, H. (2023a). *plyr: Tools for splitting, applying and combining data* [R package, Version 1.8.9]. https://CRAN.R-project.org/package=plyr

Wickham, H. (2023b). *stringr: Simple, consistent wrappers for common string operations* [R package, Version 1.5.1]. https://CRAN.R-project.org/package=stringr

Wickham, H. (2023c). *tidyverse: Easily install and load the "Tidyverse"* [R package, Version 2.0.0]. https://CRAN.R-project.org/package=tidyverse

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & van den Brand, T. (2024). *ggplot2: Create elegant data visualisations using the grammar of graphics* [R package, Version 3.5.0]. https://cran.r-project.org/package=ggplot2

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* [R package, Version 1.1.4]. https://CRAN.R-project.org/package=dplyr

Wickham, H., Vaughan, D., Girlich, M., & Ushey, K. (2023). *tidyr: Tidy messy data* [R package, Version 1.3.1]. https://CRAN.R-project.org/package=tidyr

Wood, G., Willmes, K., Nuerk, H.-C., & Fischer, M. H. (2008). On the cognitive link between space and number: A meta-analysis of the SNARC effect. *Psychology Science Quarterly*, *50*(4), 489–525.

Zebian, S. (2005). Linkages between number concepts, spatial thinking, and directionality of writing: The SNARC effect and the REVERSE SNARC effect in English and Arabic monoliterates, biliterates, and illiterate Arabic speakers. *Journal of Cognition and Culture*, *5*(1–2), 165–190. https://doi.org/10.1163/1568537054068660

Zorzi, M., Priftis, K., Meneghello, F., Marenzi, R., & Umiltà, C. (2006). The spatial representation of numerical and non-numerical sequences: Evidence from neglect. *Neuropsychologia*, *44*(7), 1061–1067. https://doi.org/10.1016/j.neuropsychologia.2005.10.025

A4.2. Supplementary materials for Study 4

These supplementary materials belong to the following publication (see Appendix A4.1.):

Roth, L.*, Huber, J. F.*, Kronenthaler, S., van Dijck, J., Cipora, K., Butz, M. V., & Nuerk, H.-C. (2025). Looks like SNARC spirit: Coexistence of short- and long-term associations between letters and space. *Quarterly Journal of Experimental Psychology, 78*(10), 2110–2132. https://doi.org/10.1177/17470218251324437

Supplementary Material for:

Looks like SNARC spirit:

Coexistence of short- and long-term associations between letters and space

Lilly Roth[1]*, Julia F. Huber[1]*, Sophia Kronenthaler[1], Jean-Philippe van Dijck[2,3], Krzysztof Cipora[4], Martin V. Butz[5], Hans-Christoph Nuerk[1,6,7]

[1] Department of Psychology, University of Tübingen, Tübingen, Germany

[2] Department of Applied Psychology, Thomas More University, Antwerp, Belgium

[3] Department of Experimental Psychology, Ghent University, Ghent, Belgium

[4] Centre for Mathematical Cognition, Loughborough University, Loughborough, UK

[5] Department of Computer Science, University of Tübingen, Tübingen, Germany

[6] LEAD Graduate School & Research Network, University of Tübingen, Tübingen, Germany

[7] German Center for Mental Health (DZPG), Berlin, Germany

* equal contribution, shared first-authorship

E-mail-addresses: lilly.roth@uni-tuebingen.de, julia.huber@psycho.uni-tuebingen.de, sophia.kronenthaler@gmail.com, jean-philippe.vandijck@thomasmore.be, k.cipora@lboro.ac.uk, martin.butz@uni-tuebingen.de, hc.nuerk@uni-tuebingen.de

Corresponding author: Prof. Dr. Hans-Christoph Nuerk

Supplementary Material A: Stimulus selection

When selecting the letters to be used as stimuli in Experiments 1, 2, and 3, and when selecting the letter sequences to be learned in Experiments 2 and 3, we took various factors into account. Specifically, (1) the length of the sequences, (2) the letters within the sequences, (3) the order of the letters, and (4) the choice of consonants versus vowels were considered as described in detail below:

(1) Length of the letter sequences

The working memory seems to have a capacity limit between three and five stimuli (Cowan, 2001, 2010). To ensure that participants could successfully repeat the letter sequences memorize the letter sequences, we chose a length of the letter sequences of four letters. Additionally, a recent study by Vivion et al. (2024) further validated that a sequence length of four items is suitable to detect spatialization.

(2) Selection of the letters within the sequences

The experiment should be carried out using a computer keyboard. On a computer keyboard, certain keys (letters) are pressed with the right or left hand, depending on the writing style (Feit et al., 2016). It is possible that experience with certain keys generates spatial response behavior. For an overview of the locations of the letters used in the present study on typical German and English computer keyboards, see Figure 1 in the main manuscript. For example, the letter O might be judged more quickly with the right hand, as the key with the letter O is regularly pressed with the right hand. In order to keep this potential effect as constant as possible, only letters were selected that are pressed exclusively with the right or left hand, regardless of writing style (Feit et al., 2016). Moreover, the letters Y and Z were excluded because they have different positions on a German and English keyboard. The letters D and K were excluded as these were to be used as response keys. From the 11 remaining consonants (C, F, J, L, M, P, Q, R, S, W, X) and five vowels (A, E, I, O, U), this results in 13200 possible letter sequences consisting of two different consonants and vowels.

(3) Selection of the order of the letters within the sequences

The order of the letters in these 13200 possible sequences can be ascending (e.g., A C E F), descending (e.g., F E C A), and random (C A F E). According to Lindemann et al. (2008), memorizing a descending number sequence can reduce the size of the SNARC effect. Ascending letter sequences make it more difficult to distinguish between a SNARC-like effect for letters and the ordinal position effect. Therefore, we chose random sequences.

(4) Considerations regarding consonants and vowels

We made sure that each letter sequence contained a consonant and a vowel that are regularly typed with the right and left hand respectively (Feit et al., 2016). There is also the possibility that participants react more quickly to consonants or vowels with the right or left hand, regardless of the letters (MARC effect; see e.g., Cipora et al., 2019). To balance out this possible effect, no letter sequences were selected that begin or end with two vowels or consonants (like [E U L W] or [R J A O]), leading to 2596 remaining possible letter sequences.

Next, the area of the alphabet covered by the letter sequences should be as large as possible (total spread), and the letter sequence should be as evenly spaced as possible (distance measure). For instance, for [P X I A] all letters were sorted according to their alphabetical order (i.e., A, I, P, X). Then, each letter was assigned a number based on the position of the letters in the alphabet (i.e., A = 1, I = 9, P = 16, X = 24). Next, the distances between the consecutive letters were calculated by the absolute value of the position difference (i.e., |1-9| = 8, |9-16| = 7, |16 - 24| = 8). From the resulting triple (i.e., 8, 7, 8), the sum of the distances (i.e., 8 + 7 + 8 = 21) can be calculated to estimate the total spread, and the sum of the differences of the distances to each other (i.e., |8-7| + |8-8| + |7-8| = 2) provides a distance measure.

We excluded all letter sequences that have neighboring letters or a distance of one. For the remaining 2156 letter sequences the correlation between the total spread and the distance measure is r = 0.6. Letter sequences with a high total spread therefore also have a high distance measure and vice versa. This means that no letter sequence with maximum spread and minimum distance measure could be selected. Instead, letter sequences as close as possible to the center of the joint distribution were selected by choosing letter sequences

with a Mahalanobis distance less than or equal to one. The selection procedure based on the Mahalanobis distance was carried out twice; in the third run, no further letter sequences with a distance less than or equal to one could be selected. The first selection step led to a reduction to 814 letter sequences, the second to a reduction to 154 letter sequences. The remaining 154 letter sequences were divided into a total of seven letter sets each in 22 different orders. The final two sets of letters were subjectively selected with two orders: [O J R A], [R A O J], [L E W U], and [U L E W]. Care was taken to ensure that the letters did not form a known word or an abbreviation in order to avoid memorization of the letter sequences by means of chunking (Gobet et al., 2001).

In Experiment 1 and 2, the following symbols were selected as ‘no-go’ stimuli: ‡, △, ≡, and ∇. When making the selection, we ensured not to use symbols with already assigned semantics (e.g., &, %). In addition, the symbols should be as similar as possible to the letters used.

References for Supplementary Material A

Cipora, K., Soltanlou, M., Reips, U.-D. & Nuerk, H.-C. (2019). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behavior Research Methods*, *51*(4), 1676–1692. https://doi.org/10.3758/s13428-019-01213-5

Cowan, N. (2001). Metatheory of storage capacity limits. *Behavioral and Brain Sciences*, *24*(1), 154–176. https://doi.org/10.1017/s0140525x0161392x

Cowan, N. (2010). The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science*, *19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Feit, A. M., Weir, D. & Oulasvirta, A. (2016). How We Type. In J. Kaye, A. Druin, C. Lampe, D. Morris & J. P. Hourcade (Eds.), *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4262–4273). ACM. https://doi.org/10.1145/2858036.2858233

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I. & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243. https://doi.org/10.1016/S1364-6613(00)01662-4

Lindemann, O., Abolafia, J. M., Pratt, J. & Bekkering, H. (2008). Coding strategies in number space: memory requirements influence spatial-numerical associations. *Quarterly Journal of Experimental Psychology,* *61*(4), 515–524. https://doi.org/10.1080/17470210701728677

Supplementary Material B

Table B1

Mean RTs of responses for 'go' trials in ms (with SDs in brackets) per letter and per experiment

Letter	Experiment 1	Experiment 2	Experiment 3
A	559.7 (140.3)	526.2 (111.6)	579.6 (139.5)
E	581.6 (127.0)	560.8 (123.6)	613.0 (145.9)
J	587.5 (138.6)	563.8 (121.1)	633.1 (147.3)
L	621.2 (131.3)	593.3 (131.2)	647.2 (153.3)
O	568.9 (132.7)	548.0 (110.1)	586.4 (146.3)
R	578.2 (136.2)	558.4 (117.9)	608.2 (155.0)
U	592.0 (132.1)	571.6 (131.2)	635.0 (156.1)
W	590.4 (128.8)	571.4 (129.2)	616.1 (142.7)

Table B2

Mean accuracies of responses for 'go' trials per letter set / sequence and per experiment

Letter set / *sequence*	Experiment 1	Experiment 2	Experiment 3
[O J R A]	0.924	0.957	0.923
[R A O J]		0.960	0.948
[L E W U]	0.942	0.959	0.937
[U L E W]		0.938	0.945

Supplementary Material C

Table C1

Literature overview for studies that have investigated the role of magnitude vs. ordinality in spatial mental representations, with WM demands referring to the WM task, and with magnitude relevance (only for number tasks), ordinality relevance (either of sequences stored in LTM or in WM), and WM salience (referring to whether a sequence stored in WM is made salient within the judgment task), ordered chronologically by publication date.

Study and experiment	Stimuli and judgment task	WM demands	Nature of task	Magnitude relevance (stored in LTM)	Ordinality relevance (stored in LTM or WM)	WM salience	Observed effect(s)
Gevers et al. (2003), Exp. 1	Months: before vs. after July	None	Respond to all trials	No magnitude	LTM order relevant (direct task), no WM order	None	SNARC-like effect (LTM)
Gevers et al. (2003), Exp. 1	Months: ends with the letter R vs. not	None	Respond to all trials	No magnitude	LTM order irrelevant (indirect task), no WM order	None	SNARC-like effect (LTM)

Gevers et al. (2003), Exp. 2	Letters: before vs. after the letter O	None	Respond to all trials	No magnitude	LTM order relevant (direct task), no WM order	None	SNARC-like effect (LTM)
Gevers et al. (2003), Exp. 2	Letters: consonant vs. vowel	None	Respond to all trials	No magnitude	LTM order irrelevant (indirect task), no WM order	None	SNARC-like effect (LTM)
Gevers et al. (2004)	Days of the week: before vs. after Wednesday	None	Respond to all trials	No magnitude	LTM order relevant (direct task), no WM order	None	SNARC-like effect (LTM)
Gevers et al. (2004)	Days of the week: contains letter R vs. not	None	Respond to all trials	No magnitude	LTM order irrelevant (indirect task), no WM order	None	SNARC-like effect (LTM)

Van Dijck & Fias (2011), Exp. 1	Number parity: odd vs. even	Sequence stored in WM	Go/no-go design: only respond to numbers from sequence stored in WM	Magnitude irrelevant (indirect task)	LTM order irrelevant (indirect task), WM order irrelevant (indirect task)	High	OPE (WM), but no evidence for SNARC effect (LTM)
Van Dijck & Fias (2011), Exp. 2	Word meaning: fruit vs. vegetable	Sequence stored in WM	Go/no-go design: only respond to numbers from sequence stored in WM	No magnitude	LTM order irrelevant (indirect task), WM order irrelevant (indirect task)	High	OPE (WM)
Ginsburg et al. (2014), Exp. 1	Number parity: odd vs. even	Sequence stored in WM	Go/no-go design: only respond to numbers from sequence stored in WM	Magnitude irrelevant (indirect task)	LTM order irrelevant (indirect task), WM order irrelevant (indirect task)	High	OPE (WM), but no evidence for SNARC effect (LTM)
Ginsburg et al. (2014), Exp. 2	Number magnitude: small vs. large	Sequence stored in WM	Go/no-go design: only respond to numbers from sequence stored in WM	Magnitude relevant (direct task)	LTM order relevant (direct task), WM order irrelevant (indirect task)	High	OPE (WM), but no evidence for SNARC effect (LTM)

Ginsburg et al. (2014), Exp. 3	Number parity: odd vs. even	Sequence stored in WM	Respond to all trials	Magnitude irrelevant (indirect task)	LTM order irrelevant (indirect task), WM order irrelevant (indirect task)	Intermediate	SNARC effect (LTM), but no evidence for OPE (WM)
Ginsburg et al. (2014), Exp. 4	Number magnitude: small vs. large	Sequence stored in WM	Respond to all trials	Magnitude relevant (direct task)	LTM order relevant (direct task), WM order irrelevant (indirect task)	Intermediate	SNARC effect (LTM), but no evidence for OPE (WM)
Guida et al. (2015)	Letter probe recognition: part of the auditory sequence vs. not	Sequence stored in WM	Respond to all trials	No magnitude LTM order irrelevant	(indirect task), WM order irrelevant (indirect task)	High	OPE (WM), but SNARC-like effect (LTM) not tested

Huber et al. (2016)	Number parity: odd vs. even	Sequence stored in WM	Go/no-go design: only respond to numbers from sequence stored in WM	Magnitude irrelevant (indirect task)	LTM order irrelevant (direct task), WM order irrelevant (indirect task)	High	Both SNARC effect (LTM) and OPE (WM)
Current study, Exp. 1	Letters: consonant vs. vowel	None	Go/no-go design: only respond to numbers	No magnitude	LTM order irrelevant (indirect task), WM order irrelevant (indirect task)	None	SAARC effect (LTM), but no evidence for OPE (WM)
Current study, Exp. 2	Letters: consonant vs. vowel	Sequence stored in WM	Go/no-go design: only respond to numbers	No magnitude	LTM order irrelevant (indirect task), WM order irrelevant (indirect task)	Intermediate	SAARC effect (LTM), but no evidence for OPE (WM)
Current study, Exp. 3	Letters: consonant vs. vowel	Sequence stored in WM	Go/no-go design: only respond to numbers from sequence stored in WM	No magnitude	LTM order irrelevant (indirect task), WM order irrelevant (indirect task)	High	Both SAARC effect (LTM) and OPE (WM)

References for Supplementary Material C

Gevers, W., Reynvoet, B., & Fias, W. (2003). The mental representation of ordinal sequences is spatially organized. *Cognition*, *87*(3), 87–95. https://doi.org/10.1016/S0010-0277(02)00234-2

Gevers, W., Reynvoet, B., & Fias, W. (2004). The mental representation of ordinal sequences is spatially organized: evidence from days of the week. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *40*(1), 171–172. https://doi.org/10.1016/S0010-9452(08)70938-9

Ginsburg, V., van Dijck, J.-P., Previtali, P., Fias, W., & Gevers, W. (2014). The impact of verbal working memory on number-space associations. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(4), 976–986. https://doi.org/10.1037/a0036378

Guida, A., Leroux, A., Lavielle-Guida, M., & Noël, Y. (2015). A SPoARC in the dark: Spatialization in verbal immediate memory. *Cognitive Science, 40*(8), 2018–2121. https://doi.org/10.1111/cogs.12316

Huber, S., Klein, E., Moeller, K., & Willmes, K. (2016). Spatial-Numerical and ordinal positional associations coexist in parallel. *Frontiers in Psychology*, 7:438, 1–13. https://doi.org/10.3389/fpsyg.2016.00438

van Dijck, J.-P., & Fias, W. (2011). A working memory account for spatial–numerical associations. *Cognition*, *119*(1), 114–119. https://doi.org/10.1016/j.cognition.2010.12.013

Supplementary Material D

Comparisons of dRTs between letters and experiments

To explore whether WM salience (which was higher in Experiment 3 than in Experiment 2) affected spatial response preferences for some letters more than for others, we ran a two-way mixed ANOVA for dRT separately for each memorized sequence. Each of the four ANOVAs considered one between-subjects factor (WM salience: intermediate in Experiment 2 vs. high in Experiment 3) and one within-subjects factor (letter: A vs. J vs. O vs. R when the memorized sequence was [O J R A] or [R A O J], and E vs. L vs. U vs. W when the memorized sequence was [L E W U] or [U L E W]). A Bonferroni-Holm correction was applied for the interpretation of the four ANOVA outcomes, starting with a significance level of $\alpha = .05 / 4 = .0125$ for the lowest p-value.

Most importantly, the interaction effect between WM salience and memorized sequence remained non-significant for each letter sequence ([O J R A]: $F(3, 138) = 1.27$, $p = .287$; [R A O J]: $F(3, 135) = 2.21$, $p = .089$; [L E W U]: $F(3, 177) = 1.91$, $p = .130$; [U L E W]: $F(3, 162) = 0.967$, $p = .410$). In other words, we did not find evidence for the stability (in terms of robustness against increasing salience of ordinality stored in WM) of spatial response preferences to be stronger for some letters than for others. Moreover, no significant main effect of WM salience was found (averaged over letters), pointing towards similar dRTs in Experiments 2 and 3 ([O J R A]: $F(1, 46) = 0.926$, $p = .341$; [R A O J]: $F(1, 45) = 0.001$, $p = .971$; [L E W U]: $F(1, 59) = 0.816$, $p = .370$; [U L E W]: $F(1, 54) = 0.541$, $p = .465$). However, spatial response preferences differed between letters in each of the four letter sequences (averaged over both Experiments 2 and 3), as revealed by a significant main effect of letter in each of the four ANOVAs ([O J R A]: $F(3, 138) = 2.68$, $p < .050$; [R A O J]: $F(3, 135) = 8.48$, $p < .001$; [L E W U]: $F(3, 177) = 4.73$, $p = .003$; [U L E W]: $F(3, 177) = 1.91$, $p = .003$).

A5. Study 5

A5.1. Publication of Study 5

Citation:
Roth, L., Caffier, J., Reips, U.-D., Nuerk, H.-C., Overlander, A. T., & Cipora, K. (2025). One and only SNARC? Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude. *Royal Society Open Science, 12*(1), Article 241585. https://doi.org/10.1098/rsos.241585

Copyright:
© 2024 Roth, Caffier, Reips, Nuerk, Overlander, Cipora (Open Access)
This article is published under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, provided the original authors and source are credited.

ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

CC BY

Registered report

Cite this article: Roth L, Caffier J, Reips U-D, Nuerk H-C, Overlander AT, Cipora K. 2025 One and only SNARC? Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude. *R. Soc. Open Sci.* **12**: 241585.
https://doi.org/10.1098/rsos.241585

Received: 29 September 2024
Accepted: 9 October 2024

Subject Category:
Psychology and Cognitive Neuroscience

Subject Areas:
Psychology

Keywords:
Spatial-Numerical Associations, SNARC effect, mental number line, replication, flexibility

Author for correspondence:
Krzysztof Cipora
e-mail: k.cipora@lboro.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.7560265.

THE ROYAL SOCIETY PUBLISHING

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

One and only SNARC? Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude

Lilly Roth[1], John Caffier[3], Ulf-Dietrich Reips[3], Hans-Christoph Nuerk[1,2,4], Annika Tave Overlander[3] and Krzysztof Cipora[5]

[1]Department of Psychology, and [2]LEAD Graduate School & Research Network, University of Tübingen, Tübingen, Germany
[3]Department of Psychology, University of Konstanz, Konstanz, Germany
[4]German Center for Mental Health (DZPG), Mannheim, Germany
[5]Centre for Mathematical Cognition, Loughborough University, Loughborough, UK

LR, 0000-0002-2690-0504; JC, 0000-0002-0475-9167; U-DR, 0000-0002-1566-4745; H-CN, 0000-0002-0331-7498; ATO, 0009-0006-8373-4086; KC, 0000-0003-0077-9336

Numbers are associated with space, but it is unclear how flexible these associations are. We investigated whether the SNARC effect (Spatial-Numerical Association of Response Codes; Dehaene *et al*. 1993 *J. Exp. Psychol*. **122**, 371–396. (doi:10.1037/0096-3445.122.3.371); i.e. faster responses to small/large number magnitude with the left/right hand, respectively) is fully flexible (depending only on relative magnitude within a stimulus set) or not (depending on absolute magnitude as well). Evidence for relative-magnitude dependency came from studies observing that numbers 4 and 5 were associated with the right in a 0–5 range but with the left in a 4–9 range (Dehaene *et al*. 1993; Fias *et al*. 1996 *Math. Cogn*. **2**, 95–110 (doi:10.1080/135467996387552). Within this Registered Report, we conducted two online experiments running Bayesian analyses with optional recruitment stopping at moderate evidence (BF_{10} above 3 or below 1/3). Experiment 1 (n = 200) replicated relative-magnitude dependency using the original stimuli. However, Experiment 2 (n = 300) additionally demonstrated absolute-

© 2024 The Author(s). Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

magnitude dependency, while considering recent advances in SNARC research using 1–5 excluding 3 and 4–8 excluding 6. In contrast to the frequently perpetuated notion of fully flexible Spatial-Numerical Associations, some fixed relation to absolute magnitude prevails. These findings have important consequences for understanding how Spatial-Numerical Associations might support numerical processing.

1. Introduction

Numbers are highly relevant in everyday life. Therefore, much research has been devoted to understanding how we process and represent them in our minds. Interestingly, various aspects of numerical information such as cardinality and ordinality are systematically associated with different aspects of space such as extensions or directions [1–3]. This broad range of phenomena is referred to under the umbrella term Spatial-Numerical Associations (SNAs) [4,5]. Investigating these associations is fundamental for models of number representation and—considering the bigger picture—of human cognition.

The hallmark directional SNA is the Spatial-Numerical Association of Response Codes (SNARC) effect, which denotes that in left-to-right reading cultures, participants respond faster to small/large magnitude numbers on the left/right side, respectively [6]. Interestingly, the SNARC effect can be observed in a parity judgement task, in which the magnitude of the numbers is not task-relevant. This effect has been replicated using different modalities, set-ups and tasks (see [7] for an online replication; [5,8] for a recent review; [9] for a meta-analysis). The SNARC effect is typically quantified using the repeated-measures regression originally proposed by Lorch & Myers [10] and applied to the SNARC effect by Fias *et al.* [8]. In the first step mean differences in reaction times (RTs) between the right and left hand (dRTs) are regressed on numerical magnitude for each participant separately. A negative slope indicates an increasing right-hand advantage with increasing number magnitude (the more negative the so-called SNARC slope, the stronger the SNARC effect). Subsequently, to check for the SNARC effect at the group level, individual SNARC slopes are tested against zero with a one-sample *t*-test.

Interestingly, several studies have documented that the SNARC effect is not fixed but might be prone to several types of manipulation ([11] for a taxonomy), for instance, changing the number range of the used stimuli, which has been classified as representational, intra-experimental manipulation. The spatial mental number representation seems to be adapted to fit the task at hand. In this work, we focus on the extent to which the SNARC effect flexibly adjusts to the specific range of the numbers being used in the task set.

1.1. Relative-magnitude (RM) dependency of the SNARC effect

The seminal paper by Dehaene *et al*. [6] has already demonstrated in its Experiment 3 that the SNARC effect depends on the relative rather than the absolute magnitude (AM) of numbers. They found the SNARC effect in two different numerical intervals ranging from 0 to 5 and from 4 to 9. In the lower interval, responses to numbers 4 and 5 were faster with the right hand than with the left (typical response pattern for large numbers) and right-hand responses to these numbers were faster than right-hand responses to lower numbers. In contrast, in the higher interval, responses to these numbers were faster with the left hand than with the right (typical response pattern for small numbers) and left-hand responses to these numbers were faster than left-hand responses to higher numbers. This finding was replicated by Fias *et al*. [8] (Experiment 1). It suggests that the SNARC effect dynamically adapts to the current task set (i.e. numbers being used) and is determined by the RM of the number within the set rather than its AM. We refer to this claim about the SNARC effect as *RM dependency*.

RM dependency is considered as one of the crucial features of the SNARC effect and has been taken for granted since these early findings. The results of Dehaene *et al*.'s [6] and Fias *et al*.'s [8] experiments are widely cited as an argument for the SNARC effect being dependent on the given number range (e.g. by [12–17]). The RM dependency of the SNARC effect has been demonstrated by several other studies even going beyond a basic set-up comprising judgements on single-digit numbers. For instance, Tlauka [18] found a SNARC effect both when using the two numbers, 1 and 100, and when using the two numbers 100 and 900. The number 100 was associated with the right/left

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

when it was the larger/smaller of the two numbers, respectively. Ben Nathan *et al*. [19] went even further, showing that the SNARC effect is not only RM dependent on the task level but built up on a trial-to-trial basis. They found the right- and left-key response speed advantages in magnitude judgement tasks to depend on the RM in comparison to the ever-changing reference number. What is more, evidence for RM dependency of SNARC-like effects goes beyond numerical stimuli. Wühr & Richter [20] found a SNARC-like effect (association of physically smaller/larger stimuli with the left/right, respectively) to depend on relative rather than absolute stimulus size.

Importantly, RM dependency has also been used as a methodological tool to show that a spatial-numerical phenomenon is in fact the SNARC effect. For instance, Rugani *et al*. [21], Di Giorgio *et al*. [22] and Giurfa *et al*. [23] showed RM dependency to claim that a certain effect they observed in newly hatched chickens, in newborn children and in honeybees is of the same nature as the SNARC effect. To sum up, there is evidence for the RM dependency of the SNARC effect in various tasks and set-ups, and it has even been used to validate SNAs.

1.2. RM dependency in the light of number-representation models

RM dependency fits well with most theoretical accounts of number representation. The seminal work of Restle [24] outlining the mental number line (MNL) account, which has been proposed as the first explanation for the SNARC effect [6], postulates that the MNL is flexible and dynamically adapts to the task demands. In line with this, Pinhas *et al*. [17] claim that the resolution of the MNL can be adjusted to the numerical context. The accounts of verbal–spatial coding [25] and polarity correspondence [26] are on the one hand in line with RM dependency, but on the other hand, they do not make clear statements about RM being the *only* decisive factor determining the SNARC effect. Crucially, both accounts assume that long-term number representations underlie the SNARC effect, which hardly justifies the SNARC effect's flexibility [15,27]. The working memory account [28,29] originally claimed that the SNARC effect does not rely on long-term number representations but is instead constructed during task execution, which speaks in favour of pure RM dependency. However, Ginsburg *et al*. [14] and Koch *et al*. [30] argue that short-term number representations do not always fully overrule long-term number representations. This idea has been incorporated in the hybrid account proposed by van Dijck *et al*. [27] as well, and it allows the coexistence of RM dependency and dependency of the SNARC effect on AM (henceforth *AM dependency*). Furthermore, concurrent RM dependency and AM dependency would also be in line with the idea that multiple number representations and multiple spatial reference frames can be activated and operated simultaneously [31]. To conclude, the assumption that AM plays no role can hardly be derived from theoretical accounts of the SNARC effect.

1.3. Hints towards AM dependency of the SNARC effect

In addition to the prominent claims on the RM dependency of the SNARC effect, the literature also provides hints towards an AM dependency of the SNARC effect. It is important to note that AM dependency can, on the one hand, influence the strength of the SNARC effect (reflected by the SNARC slope), and on the other, the location of numbers on the MNL in absolute terms (reflected by the intercept of the regression line and by dRTs of critical numbers that are part of both number ranges). Crucially, the SNARC effect seemed to be stronger in the lower than in the higher number range in both initial studies demonstrating the RM dependency (−20.1 ms versus −10.9 ms in Dehaene *et al*. [6]; and −10.18 versus −7.19 ms in Fias *et al*. [8]), suggesting AM dependency as well. In Fias *et al*.'s (1996) results, the observed slope difference had approximately an effect size of Cohen's $d = 0.16$ (i.e. the slope difference of 2.99 divided by the standard deviation [SD] for this slope difference of 18.34 ms, which has been calculated with SD = 15.1 ms and SD = 11.2 ms for the lower and higher number ranges, assuming a rather conservative correlation between them of $r = 0.05$, which corresponds to the correlation we have observed in our previous colour judgement tasks, see Roth *et al*. [32], where we also found a stronger SNARC effect in the lower than in the higher half of the stimulus set ranging from 1 to 9). Moreover, the results pointed towards an overall shift of small/large numbers to the left/right on the MNL, respectively, since the smallest-number intercept (i.e. the predicted dRT for the smallest number magnitude of the range, which was 0/4 in the lower/higher range, respectively) was larger in the lower than in the higher range (37.52 ms versus 14.03 ms in Dehaene *et al*. [6], and 15.43 ms versus 8.82 ms in Fias *et al*. [8]). However, the mean-number intercepts (i.e. the predicted dRT for

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

the mean number magnitude of the range, which was 2.5/6.5 in the lower/higher range, respectively) did not differ much in Fias *et al*.'s results (−10.02 ms versus −9.16 ms). In Dehaene *et al*.'s results, this intercept seemed to be smaller in the higher number range, but it cannot be calculated exactly based on the data reported in the paper.

1.4. Methodological limitations of the two initial studies demonstrating RM dependency

Even if we use the two original studies as guidance for further investigations, their findings are not reliable because of several important limitations regarding the design and the interpretation of the results. Both Dehaene *et al*. [6] and Fias *et al*. [8] found a significant two-way interaction of response side (left versus right) and magnitude (small versus medium versus large). Apart from the repeated-measures regression approach, the SNARC effect can also be quantified as a two-way interaction of response side and magnitude (for methodological considerations, see Fias *et al*. [8]) or as linear contrast in an ANOVA [33]. However, the three-way interaction of response side and magnitude with interval (0–5 versus 4–9) remained non-significant in both studies. In Fias *et al*.'s [8] additional repeated-measures regression, the resulting SNARC slopes differed significantly from zero in both intervals in a one-sample *t*-test, and the difference in SNARC slopes between both intervals remained non-significant in a paired *t*-test. Crucially, the strong conclusion of pure RM dependency that has been derived from these null results is dangerously close to mistaking the absence of evidence for evidence of absence. Importantly, no Bayesian analysis was conducted to test whether the null results supported the null hypothesis (and it is not possible to run a post-hoc Bayesian analysis due to the lack of a report of the exact *t*-statistic). What is more, neither Dehaene *et al*. [6] nor Fias *et al*. [8] tested whether the dRT pattern for the same number differed significantly between number ranges—even if the right-hand advantage (reflected by negative dRTs) for numbers 4 and 5 in the range from 0 to 5 and the left-hand advantage (reflected by positive dRTs) for these numbers in the range from 4 to 9 are often cited. Also, the smallest-number intercepts and the mean-number intercepts were not compared between ranges.

Moreover, the design was most likely underpowered for the relevant statistical comparisons in both studies (see below for calculations). On the one hand, this was due to the relatively low sample sizes ($n = 12$ in Dehaene *et al*. [6] and $n = 24$ in Fias *et al*. [8]). On the other, only 15 repetitions per experimental cell (i.e. per number magnitude and response-key assignment) were used. Later methodological studies proposed to use at least 20 repetitions and 20 participants to detect the SNARC effect, and even more repetitions and participants to detect differences in the size of the SNARC effect [34]. Following the *effect-size sensitivity approach* [35], we have run power calculations to determine SNARC slope differences between the two number ranges that are detectable in a *t*-test for two dependent samples at different adequate power levels (adapting Monte-Carlo simulations by Wickelmaier [36]). For the sample size used by Fias *et al*. [8] and with the SD they observed, our calculations revealed that at power levels of 0.80, 0.90 and 0.95, only SNARC slope differences between the two number ranges of minimum 11.0 ($d = 0.60$), 12.7 ($d = 0.69$) and 14.1 ms ($d = 0.77$) could have been detected, respectively. Note that we ran these calculations within the frequentist framework, which corresponds to the data analysis by Fias *et al*. (for calculations in both the frequentist and the Bayesian framework, see 'PCI Registered Report Materials' at [37], created using the R packages *rmarkdown* by Allaire *et al*. [38]; *knitr* by Xie [39], and *BayesFactor* by Morey *et al*. [40]). However, such differences in SNARC slopes are very unlikely, even in the case of AM dependency, because they would be larger than the typically observed SNARC slopes themselves. Because of the lack of related information in Dehaene *et al*.'s (1993) paper, we were not able to run such power calculations for their results; but because their sample was even smaller, they could have detected only even larger differences.

Moreover, the stimuli used in both studies (0, 1, 2, 3, 4, 5 and 4, 5, 6, 7, 8, 9) lead to two problems. First, the average number magnitude in both number ranges is larger for odd than for even numbers (3 versus 2 in the lower and 7 versus 6 in the higher number range). This can lead to a confound with the MARC (Linguistic Markedness of Response Codes) effect that denotes a left-/right-hand advantage when responding to odd/even numbers, respectively [41]. Such a confound may decrease the SNARC effect [33,42]. The association of small/large numbers to the left/right side, respectively, should be weaker if small/large numbers are more often even/odd, respectively. More recent studies have addressed this issue by using stimuli sets in which number magnitude and contrast-coded parity are orthogonal (e.g. [7]). Typically, it is done by using the number set 1, 2, 3, 4, 6, 7, 8 and 9, which importantly also excludes 0.

Second, using the number 0 is problematic due to its special status shown in several studies: reading time for 0 is significantly longer than for any other single-digit number and is not predicted

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

by factors determining the reading time of other single-digit numbers [43]. Nuerk *et al*. [41] and Nieder [44] provide further empirical evidence that 0 may not be represented on the MNL along with other numbers (but see Pinhas & Tzelgov [45], for another conclusion). Additionally, quite often participants have problems understanding the parity status of 0 [46]. Using 0 also turned out problematic in SNARC studies: the RTs and dRTs for the number 0 do not strongly correlate with the RTs and dRTs of other even numbers [41]. Later studies on the SNARC effect excluded 0 from the stimuli set (e.g. [13,25,47–49]). Ultimately, both the parity status and the presence of 0 might have confounded the results of the previous studies (see table 1). Therefore, in addition to the replication that we conducted as close as possible to the original studies by Dehaene *et al*. [6] and Fias *et al*. [8], we also ran a conceptual replication using suitable stimulus sets to disentangle these potential confounds and tackle all the above-mentioned limitations.

1.5. The SNARC effect operating on two reference frames at once

As we laid out so far, there is a general tendency to interpret the SNARC effect as entirely flexible based on the findings of RM dependency and on the inference-statistical null effects concerning AM dependency (in underpowered studies). However, the SNARC effect could be operating concurrently in both relative and absolute terms. Indeed, one of us has proposed that the SNARC effect operates on multiple number lines [31]. However, that paper is not about whether the SNARC effect operates on multiple number lines in terms of RM dependency and AM dependency, but instead, it used two-digit numbers as stimuli to see whether separate number lines are activated for decade and unit numbers. The operations on different number ranges are for decade and unit digits of one two-digit number (i.e. the same number, but different digits of its decomposition). Thus, the paper by Weis *et al*. provides the principal account that the SNARC effect could operate on multiple reference frames at once. The current study goes beyond their findings because it seeks to demonstrate that both RM-dependent and AM-dependent spatial mappings are concurrently present in the same digit.

1.6. The current study

In this study, we aim to answer the question whether the SNARC effect depends only on RM or whether AM plays a role as well. Crucially, in contrast to previous literature about the flexibility of the SNARC effect, we differentiate between two concepts that can be affected by RM dependency and AM dependency:

(i) On the one hand, the number mapping on the MNL (e.g. dRT for number 4) may be different depending on the experimental set-up. In our set-up, it can be RM-dependent (i.e. depending on the position on the used range, e.g. position 5 for range 0–5, or 1 for range 4–9), AM dependent (i.e. depending on the magnitude, e.g. 4), or both at the same time.
(ii) On the other hand, the strength of the SNARC effect relies on the relative increase of right-hand advantage per increase in magnitude (i.e. the steepness of the SNARC slopes, e.g. −5 ms per number or −10 ms per number) and these slopes can differ between ranges.

For a more detailed and complex elaboration of six possible scenarios combining different parameters of (i) and (ii), see electronic supplementary material, figures S1 and S2 (see 'PCI Stage 2 Registered Report Materials' at [37].

To answer the research question, we first replicated Experiment 3 by Dehaene *et al*. [6], which has also been replicated in Experiment 1 by Fias *et al*. [8], using the original number ranges from 0 to 5 and from 4 to 9. Second, we conducted a conceptual replication to address confounds due to the unequal distribution of odd and even numbers and due to the presence of 0 in both stimuli sets, using the number ranges 1 to 5 (excluding 3) and 4 to 8 (excluding 6). The middle number of the range is also excluded in most SNARC studies using the typical set from 1 to 9. This way, the critical numbers that appear in both ranges were the same in both experiments, namely 4 and 5. Table 1 gives an overview of the used number ranges and of confounds between number parity and number magnitude in Experiment 1 that were avoided in Experiment 2.

In both of our replication experiments, high statistical power was obtained by testing much larger samples than Dehaene *et al*. [6] and Fias *et al*. [8] and by increasing the number of repetitions per experimental cell from 15 to 25. To be able to quantify evidence both for differences between number

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

Table 1. Stimulus sets and their characteristics.

Experiment 1 (close replication: number ranges used by [6,8])				Experiment 2 (conceptual replication)			
lower range		higher range		lower range		higher range	
absolute magnitude predictor	contrast-coded parity predictor	absolute magnitude predictor	contrast-coded parity predictor	absolute magnitude predictor	contrast-coded parity predictor	absolute magnitude predictor	contrast-coded parity predictor
0	+0.5	**4**	+0.5	1	−0.5	**4**	+0.5
1	−0.5	**5**	−0.5	2	+0.5	**5**	−0.5
2	+0.5	6	+0.5	**4**	+0.5	7	−0.5
3	−0.5	7	−0.5	**5**	−0.5	8	+0.5
4	+0.5	8	+0.5	—	—	—	—
5	−0.5	9	−0.5	—	—	—	—
mean number magnitude depending on number parity:							
$M_{even} = 2$		$M_{even} = 6$		$M_{even} = 3$		$M_{even} = 6$	
$M_{odd} = 3$		$M_{odd} = 7$		$M_{odd} = 3$		$M_{odd} = 6$	
correlation between number magnitude and number parity:							
$r = -0.293$				$r = 0$			

This table lists all stimuli used in the two experiments. It shows the confound between number parity and number magnitude in both number ranges of Experiment 1 and illustrates how we avoided it in both number ranges of Experiment 2, such that number parity and number magnitude were uncorrelated (i.e. they were orthogonal to each other as predictors in regression models). Number parity was contrast-coded with −0.5 for odd and +0.5 for even numbers when measuring the MARC effect. The number 0 was included in Experiment 1, but not used in the conceptual replication in Experiment 2 because of its special features and irregular mental representation (as outlined in the Introduction). The numbers 4 and 5, which are written in bold in the table, were present in each of the number ranges.

ranges and lack thereof, we applied the Bayesian instead of frequentist approach in statistical analysis. For the interpretation of different values for the Bayes factors, we followed the recommendations by Dienes [50]: a BF_{10} greater than 3 or 10 was treated as moderate or strong evidence for the alternative hypothesis, while a BF_{10} smaller than 1/3 or 1/10 was treated as moderate or strong evidence for the null hypothesis, respectively.

Online experiments offer the possibility to collect data from large samples and therefore reach high statistical power [51,52]. The SNARC effect has been successfully replicated in online settings (e.g. [7,30,49]). The measurement in the online set-up showed a similar reliability and magnitude compared to the SNARC effect that is typically observed in lab studies. Further, it seems to be valid as regards the correlations of the SNARC effect with mean RTs and SD of RTs, which are similar to lab studies.

In this study, we expected to replicate the findings by Dehaene *et al*. [6] and by Fias *et al*. [8] as concerns RM dependency. However, we also expected to find evidence towards AM dependency of the number mapping on the MNL and of the strength of the SNARC effect. Previous studies have indicated tendencies that cannot be explained by RM dependency alone. Thus, we hypothesized the following:

(1) A SNARC effect in both (a) the lower and (b) the higher number ranges in each experiment. (a) The SNARC effect in the lower range served as a manipulation check and was considered as a prerequisite for testing hypotheses 2 and 3 in each experiment. Both (a) and (b) aimed at replicating the results by Dehaene *et al*. [6] and Fias *et al*. [8].
(2) Both (a) RM dependency and (b) AM dependency of the number mapping on the MNL, such that small/large numbers in relative and absolute terms are shifted towards the left/right, respectively. (a) RM dependency is reflected by dRTs for the same critical numbers (i.e. 4 and 5) differing between ranges, showing that the MNL adapts flexibly and relative to the range. (b)

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

AM dependency is reflected by dRTs for these critical numbers being equal between ranges, and by dRTs for the smallest number (Experiment 1: 0 in the 0–5 range versus 4 in the 4–9 range; Experiment 2: 1 in the 1–5 range [excluding 3] versus 4 in the 4–8 range [excluding 6]) differing between ranges. AM dependency means that small/large numbers are shifted to the left/right on the MNL, although they are in exactly the same position within their range, but differ in terms of AM.

(3) AM dependency of the strength of the SNARC effect, such that it is stronger in the lower than in the higher ranges. This is reflected by steeper (i.e. more negative) SNARC slopes in the lower than in the higher ranges, which was descriptively observed in the two seminal studies by Dehaene *et al*. [6] and Fias *et al*. [8].

2. Materials and method

This study has been approved by the ethics committee of the University of Tübingen's Department of Psychology. The Stage-1 Registered Report received in-principle acceptance by the Peer Community In (PCI) on 3 December 2023, and the Stage-2 Registered Report received their final recommendation on 6 September 2024 (https://doi.org/10.24072/pci.rr.100808).

2.1. Sample size considerations

For this study, we defined Cohen's d = 0.15 as the minimal effect size of interest, because the most crucial aim of the present study was to find out whether AM dependency of the strength of the SNARC effect exists or not (hypothesis 3). By choosing this minimal effect size of interest, we were able to find evidence for or against the SNARC slope differences between number ranges that had been descriptively reported in the original studies that we wished to replicate. Due to the lack of a report of SD, it was not possible to calculate Cohen's d for the slope difference of 9.2 ms found by Dehaene *et al*. [6], but the slope difference of 2.99 ms with its SD of 18.34 ms found by Fias *et al*. [8] corresponds to an effect size of d = 0.16. Note that in the two original studies, the symmetric confidence intervals for these estimates must also include at least the double slope difference and effect size due to their non-significance. Hence, in case of AM dependency of the strength of the SNARC effect, the true effect size might in fact be larger than d = 0.15. This sample size estimation was also valid for testing hypotheses 1 and 2, which required one-sample t-tests. The reason was that an effect smaller than d = 0.15 would not be meaningful for the SNARC effect in the lower (hypothesis 1a) or higher (hypothesis 1b) number range, or for RM dependency (hypothesis 2a) and AM dependency (hypothesis 2b) of the number mapping on the MNL either. Similarly, the chosen maximal sample size was large enough to find at least moderate evidence in case these hypotheses are false.

To ensure a probability of 0.90 for finding at least moderate evidence for a true underlying effect (i.e. BF_{10} above 3, according to Dienes [50] with a minimally relevant effect size of Cohen's d = 0.15 in one-sample or paired t-tests, the sample needed to consist of 800 participants in each experiment (for power calculations, see 'PCI Registered Report Materials' at [37]. The sample size of 800 participants was required for a proportion of at least 0.90 Bayesian t-tests to yield a BF_{10} above 3 when 5000 samples of SNARC slope differences randomly drawn from a normal distribution around the minimally relevant effect size of d = 0.15 were simulated (for a similar approach, see [53]). Following the same procedure, we found that the sample needed to consist of 180 participants to ensure a probability of 0.90 for finding at least moderate evidence against an effect that is truly absent (i.e. BF_{10} below 1/3 for d = 0, according to Dienes [50]). Note that the sample size of 180 was smaller than the initial sample size of 200 that was collected in the 'Sequential Bayes Factor with maximal n' (SBF + maxN) approach as described by Schönbrodt & Wagenmakers [54] (see explanation below). For these calculations, we used SD = 15.1 ms and SD = 11.2 ms for the lower and higher number ranges, as reported by Fias *et al*. [8], although the SD in our previous colour judgement experiments were only SD = 4.2 ms and SD = 3.9 ms [32]. Hence, our calculations were rather conservative, and the probability of finding evidence for a true underlying effect thus was most probably even higher. While in the frequentist framework, low error type II rates (and high statistical power) need to be achieved, in the Bayesian framework, a low probability of misleading evidence for the null hypothesis in case of a true underlying effect and a high probability of finding evidence for a true underlying effect needs to be ensured. To achieve the same probability for error type II and misleading evidence, Bayesian

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

t-tests (using the default *r*-scale of 0.707 as uninformed prior in the Cauchy distribution) require larger samples as compared to frequentist *t*-tests [53].

Importantly, as we ran Bayesian instead of frequentist analyses, we made use of the SBF + maxN approach and defined an optional stopping threshold to make our data collection more efficient. Namely, we used moderate evidence in favour of all hypotheses ($BF_{10} > 3$) or against them ($BF_{10} < 1/3$) as thresholds. More precisely, for each experiment, we first recruited 200 participants (i.e. complete individual datasets) and computed the BF_{10} for the SNARC effect in lower (hypothesis 1a) and higher (hypothesis 1b) number ranges, for the shift of critical small/large numbers in both relative (hypothesis 2a) and absolute (hypothesis 2b) terms towards the left/right, respectively, and for the SNARC slope difference between ranges (hypothesis 3). As long as the BF_{10} did not reach any of the two thresholds for all hypotheses, we collected another 20 complete individual datasets and recalculated the BF_{10}. If no threshold had been reached with our maximal sample size of 800 participants (that is required for obtaining at least moderate evidence for a true underlying minimally relevant effect with a probability of at least 0.90, as explained above), we would have stopped the sequential recruiting of participants in any case.

2.2. Participants

For each experiment, adults were recruited via the recruiting platform Prolific. To comply with our ethics proposal, they had to be at least 18 years old, and because of possible age differences in RTs, we set the maximum age to 40 years. As the experiments were conducted in English, participation was only possible for native English speakers (as per Prolific's screening based on self-reports). Participation took approximately 20 min and was compensated with £5 (partial payment for partial participation).

2.3. Design and experimental task

In the parity judgement task with binary response-key set-up, participants had to indicate as fast and as accurately as possible whether the number presented on the screen was odd or even. The parity judgement task is widely used in numerical cognition and the standard task to investigate the SNARC effect (see [5] for a review and Wood et al., 2008 [9] for a meta-analysis). We assigned participants randomly to one of our two experiments. In Experiment 1 (close replication of Dehaene *et al*. and Fias *et al*. [6,8]), the numbers from 0 to 5 were used in the lower number range and the numbers from 4 to 9 in the higher number range. In Experiment 2 (conceptual replication), the numbers from 1 to 5 (excluding 3) were used in the lower number range and the numbers from 4 to 8 (excluding 6) in the higher number range, eliminating confounds between number parity and number magnitude (see table 1) and special influences of the number 0.

In both experiments, we used 25 repetitions per number magnitude in each number range (lower versus higher) and each response-key assignment (MARC congruent, i.e. left-hand responses to odd and right-hand responses to even numbers, versus MARC incongruent, i.e. right-hand responses to odd and left-hand responses to even numbers). This lead to a total of 600 trials for Experiment 1 and 400 trials for Experiment 2. In each experiment, the trials were equally divided into four blocks (one per combination of number range and response-key assignment), with a break of minimum 30 s between them. Participants were randomly assigned to one of four block orders of congruency and number range (see figure 1). The order of stimulus presentation within blocks was fully randomized. Each trial started with a square (extended ASCII 254 with a font size 72px) as an eye fixation point (300 ms). Then the number (Open Sans font, size 72px) was presented until a response was given. A blank screen (500 ms) concluded the trial. Stimuli, as well as fixation squares, were presented in black (0, 0, 0 in RGB notation), while the background remained grey (150, 150, 150 in RGB notation). The time course of an exemplary trial is illustrated in figure 2. Each block was preceded by a short practice session, in which each number was presented twice (i.e. 12 practice trials before each block in Experiment 1 and eight practice trials before each block in Experiment 2). Accuracy feedback appeared during practice sessions only.

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

	Block order 1	Block order 2	Block order 3	Block order 4
Block 1	lower range MARC incongruent	lower range MARC congruent	higher range MARC incongruent	higher range MARC congruent
Block 2	lower range MARC congruent	lower range MARC incongruent	higher range MARC congruent	higher range MARC incongruent
Block 3	higher range MARC incongruent	higher range MARC congruent	lower range MARC incongruent	lower range MARC congruent
Block 4	higher range MARC congruent	higher range MARC incongruent	lower range MARC congruent	lower range MARC incongruent

Figure 1. Counterbalancing block orders in experiments 1 and 2. *Note*. This figure shows the four block orders resulting from the combination of range (lower range versus higher range) and response-key assignment (MARC congruent, i.e. odd-left and even-right, versus MARC incongruent, i.e. even-left and odd-right). Each block was preceded by two repetitions per number as practice trials (12 trials for Experiment 1 and eight trials in Experiment 2), consisted of 25 repetitions per number as experimental trials (150 trials for Experiment 1 and 100 trials in Experiment 2) and was followed by a break.

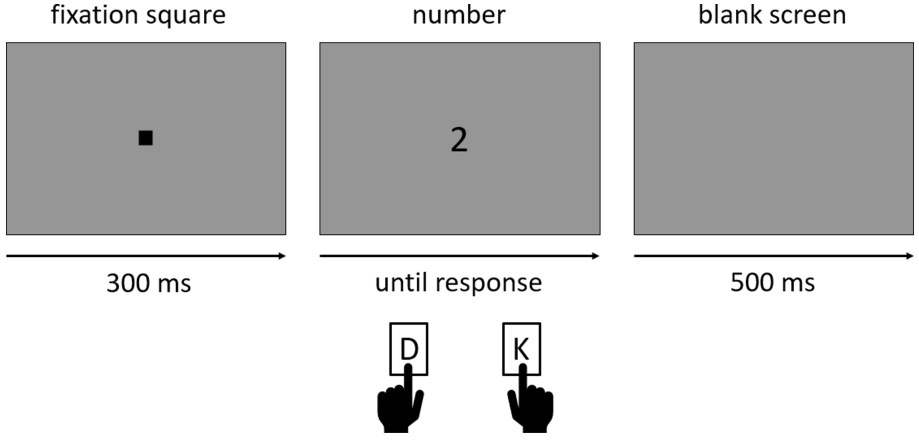


Figure 2. Time course of an exemplary trial.

2.4. Procedure

The experiments were set-up with WEXTOR (https://wextor.eu; [55]) in its HTML and JavaScript framework and adapted (see demo versions for Experiment 1 at https://exp.wextor.eu/esnarc/flex/?demo&e1 and for Experiment 2 at https://exp.wextor.eu/esnarc/flex/?demo&e2). Our previous experiments [32] have demonstrated that this software is suitable for detecting the SNARC effect in an online set-up. At the very beginning of the experiment, a seriousness check [56,57,58] was applied and participants were asked whether they wanted to participate seriously. Participants were asked to take part only if they gave their informed consent if they were using a desktop computer or laptop and if they were between 18 and 40 years old. Then, participants were asked to provide basic demographic data such as age, gender (*man*, *woman*, *other*), first native language (English and potentially others), handedness (*right-handed*, *left-handed*, *ambidextrous*) and finger-counting habits (starting hand: *left hand*, *right hand*, *does not know or no preference*; and stability: *always*, *usually*, *does not know or no preference*). For each of the above-mentioned questions, we also provided the option 'I prefer not to answer' to respect some participants' unwillingness to share information with us and to not force them to choose any option that might not reflect the truth [59,60]. Note that in earlier studies, only very few participants chose this option in any of the above-mentioned questions. Next, if not already the case for the default response keys D and K, participants could choose response keys for the experimental task which were to be located in the same row and about one hand width apart from each other on their keyboard. Then, instructions were displayed before the first block of the experimental task started with its practice trials. For instance, the instructions were as follows for the block with the lower number range in Experiment 1 (only numbers and response-to-key assignments are replaced for the higher number range or for Experiment 2): 'In our experiment, your task is to distinguish the parity of numbers, that is, to decide whether a number is even or odd. For this, please place the index finger of

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

your left hand on the [D] key and the index finger of your right hand on the [K] key on your keyboard. In each run, a black square will appear in the centre of the screen. Please look at this square. It will soon be replaced by either an even or an odd number. If the number is even (0, 2, 4), press [D]. If the number is odd (1, 3, 5), press [K]. Please answer as quickly and as accurately as possible'.

After completion of the whole experimental task, participants were asked to self-rate their math skills compared to people of their age on a visual analogue scale from *very bad* to *very good* operationalized as 0 to 400 [61]. Next, data quality was assessed by asking participants how they would describe their environment during participation (*silent*, *very quiet*, *fairly quiet*, *fairly noisy*, *very noisy* or *extremely noisy)*, whether there were any major distractions during participation (*none*, *one* or *multiple*) and whether there were any difficulties during participation (*yes* or *no*, text field for comments). Moreover, we asked participants whether they had used their left index finger for the left response key and their right index finger for the right response key throughout the experiment (*yes*, *partly* or *no*). Participants were provided with a completion code for Prolific and contact information of our research team. To prevent search engine bots (e.g. Googlebot) from submitting data on our experiment, we equipped the experiment materials with a standardized 'noindex, nofollow' meta tag, which prompts search engine bots not to index the experiment pages and also not to visit subsequent pages [62, p. 379]. Further, we restricted participation to devices over 600-pixel screen width (i.e. no smartphones). In addition, to exclude multiple submissions by the same participants, submissions from the same IP addresses were not permitted.

2.5. Data preprocessing

We used the same analysis pipeline as in another of our studies, except for not applying any colour vision check [32]. This pipeline is similar to that used by Cipora, van Dijck, *et al*. [47] in an extensive re-analysis of 18 datasets and allowed us to reliably detect the SNARC effect. The preprocessing steps were applied in the exact order as they are listed in the following. Specifically, only datasets of participants who were between 18 and 40 years old and able to seriously participate were included. Individual datasets were excluded if participants described their environment as very/extremely noisy, if they reported multiple major distractions or if they reported that they were not using their left/right index finger for the left/right response key, respectively. As outlined by Reips [52] and Birnbaum [63], experimenters were recommended to control for potential multiple submissions from the same participants by checking their User-Agents (OS and browser information) and IP addresses.[1] Regarding the data from all remaining participants, practice trials and incorrectly answered trials were not analysed. Only trials with RTs of minimum 200 ms were included in the analysis, because parity judgements faster than 200 ms are very unlikely and faster responses can therefore be treated as anticipations. Moreover, only trials with RTs of maximum of 1500 ms were included, because healthy educated adults should be capable of judging the parity status of single-digit numbers in less than 1500 ms, so that slower responses are unlikely to reflect only the mental process underlying parity judgement but instead might be caused by distractions. In the next step, further outliers were removed in an iterative trimming procedure for each participant separately, such that only RTs that are maximum 3 SD above or below the individual mean RT of all remaining trials were considered. This procedure allowed us to exclude RTs that were unlikely for each given participant and accounts for the right-skewed distribution of RTs, where the means would otherwise have been largely overestimated. After these exclusions at the trial level, only data of participants with at least 75% valid remaining trials (after excluding errors and outlier RTs) were included in the analysis at the participant level. Finally, only datasets of participants without any empty experimental cell (number magnitude per response side) in both number ranges were considered, because an empty cell would have caused a missing dRT, which in turn would have made the calculation of the SNARC slope problematic. Only complete individual datasets were included in the analysis (and none of the incomplete individual datasets fulfilled the inclusion criteria listed above).

2.6. Data analysis

All data analyses were performed in the statistical computing software R (version 4.3.3; [64]), using the R packages *BayesFactor* [40] *data.table* [65], *dplyr* [66], *GeneNet* [67], *ggplot2* [68], *neatStats* [69], *plyr* [70] and *tidyr* [71]. An overview of all research questions with corresponding hypotheses, the targeted sample size and planned analyses with a rationale, as well as the interpretations of potential outcomes

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

and theoretical conclusions, is given in the Study Design Table (see table 2). Instead of frequentist analysis, we decided to take the Bayesian approach. For this, we determined the BF_{10} associated with the corresponding Bayesian *t*-test to obtain evidence for both null and alternative hypotheses (using a default *r*-scale of 0.707 as uninformed prior using Cauchy distribution). More specifically, we calculated Bayesian *t*-tests and extracted the respective BF_{10}. Considering a BF_{10} larger than 3 as evidence against the null hypothesis is less likely than rejecting a null hypothesis with a conventional significance level of $\alpha = 0.05$ in the frequentist approach [72]. As explained above, we applied the SBF + maxN approach for sequential data analysis with optional stopping in case of at least moderate evidence for or against all hypotheses.

The key dependent variable was the mean difference between RTs of the right hand minus left hand (dRT), which was calculated for each number separately per participant and per number range. RTs were measured as the time from the onset of the number presentation on the screen until the participant's response. A potential SNARC effect can be determined by regressing dRTs on the number magnitude [8]. For each participant and for each number range, one regression was calculated. Our first dependent measure was SNARC slopes resulting from the regression of dRTs on number magnitude, which represent the change in relative advantage of right-hand compared to left-hand responses in ms per increase by one in the number magnitude (the more negative the slope, the stronger the SNARC effect). Moreover, we calculated smallest-number intercepts (when RM of the numbers in both ranges was matched, i.e. predicted dRTs for 0 in the 0–5 range versus 4 in the 4–9 range in Experiment 1, and 1 in the 1–5 range (excluding 3) versus 4 in the 4–8 range (excluding 6) in Experiment 2), as well as dRTs for critical numbers that were part of both number ranges (i.e. 4 and 5). An overview of how the tests described in the following helped us distinguish the six scenarios with different number representation shapes, depending on the number mapping on the MNL and the strength of the SNARC effect, is given in electronic supplementary material, figures S1, S2 and table S1.

First, we tested the presence of the SNARC effect on group level in both number ranges separately in each experiment (hypothesis 1). As described in the Introduction, the SNARC effect seems to be stronger in the lower than in the higher number range, resulting in a more negative slope. As the SNARC effect is very robust especially for lower ranges and possibly stronger than in higher ranges (see hypothesis 3), the SNARC effect in lower ranges (hypothesis 1a) was a manipulation check and prerequisite for following investigations (hypotheses 1b, 2 and 3). The obtained SNARC slopes were tested against zero with two-sided Bayesian one-sample *t*-tests in each number range in each experiment. This procedure corresponds to the repeated-measures regressions described by Lorch & Myers [10] and applied to the SNARC effect by Fias *et al*. [8]. It accounts for the within-subject design, where each participant completes trials for each digit in each response-to-key assignment. Although we did not expect the SNARC effect to be reversed at the group level, we preregistered two-sided tests here to stay consistent within this study. Evidence for the SNARC effect in all ranges would replicate findings from the two studies by Dehaene *et al*. [6] and Fias *et al*. [8]. The lack of conclusive evidence as regards the SNARC effect in the lower ranges (hypothesis 1a) with our maximal sample of 800 participants or even evidence against it was highly unlikely in our view. Evidence against the SNARC effect in the higher ranges (hypothesis 1b) combined with evidence for the SNARC effect in the lower ranges (hypothesis 1a) would provide support for AM dependency of the strength of the SNARC effect (hypothesis 3).

Second, to investigate RM dependency of the number mapping on the MNL, we tested whether dRTs for critical numbers (i.e. 4 and 5) differed between the lower and the higher number range (hypothesis 2a) with one two-sided paired Bayesian *t*-test per number in each experiment. Evidence for a difference would imply that the SNARC effect and the MNL are (at least partly) flexible and adapt to the number range used in a task (as in scenarios 1, 2, 4 and 5 in electronic supplementary material, figures S1 and S2). This would be in line with the literature claiming that numbers 4 and 5 are associated with the right side in the number range from 0 to 5 and with the left side in the number range from 4 to 9. However, this finding would not fully rule out AM dependency. Evidence against a difference would indicate that the SNARC effect and the MNL are AM dependent at least to some degree (as in scenarios 3 and 6 in electronic supplementary material, figures S1 and S2).

Next, to test AM dependency of the number mapping on the MNL, we tested whether the smallest-number intercepts differed between the lower and the higher number range (hypothesis 2b) with one two-sided paired Bayesian *t*-test in each experiment. Evidence for a difference would lead to the conclusion that small/large numbers are overall shifted to the left/right on the MNL, respectively (as in scenarios 2, 3, 5 and 6 in electronic supplementary material, figures S1 and S2). In other words, this would imply that the SNARC effect and the MNL are not fully RM-dependent. Evidence against a

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

difference would indicate that the SNARC effect and the MNL are at least partly RM-dependent (as in scenarios 1 and 4 in electronic supplementary material, figures S1 and S2).

Third, to investigate AM dependency of the strength of the SNARC effect, we compared SNARC slopes between the number ranges (hypothesis 3) with one two-sided paired Bayesian *t*-test in each experiment. Evidence for steeper SNARC slopes in the lower than in the higher number range can be interpreted as stronger SNARC effect within (in absolute terms) smaller than larger numbers (as in scenarios 4, 5 and 6 in electronic supplementary material, figures S1 and S2). This result would lead to the conclusion that the spatial mental representation seems to be more pronounced for small than for large numbers. Evidence against such a difference would indicate that the strength of the SNARC effect does not differ between number ranges (as in scenarios 1, 2 and 3 in electronic supplementary material, figures S1 and S2). Once the data were collected, results could be interpreted with the help of electronic supplementary material, table S1 to see which scenario most likely underlies the mental representation of number magnitude.

Additionally, to all effects reported in the unit of interest, we provide effect sizes in terms of Cohen's *d* for all Bayesian *t*-tests. Effects of $d \geq 0.2$, $d \geq 0.5$ or $d \geq 0.8$ will be interpreted as small, medium or large effect sizes, respectively.

2.7. Manipulation checks

To control the data quality in our study, we implemented a seriousness check [56,57,58] as well as a self-assessment of noise, distractions and other difficulties. To make sure that we only analysed trials that reflected mental processes in correctly executed parity judgement, we excluded incorrectly answered trials and trimmed RTs (as described in the data preprocessing pipeline in our Stage-1 Registered Report). Also, we excluded data of participants with less than 75% valid trials to only build our results on participants who understood and followed the task instructions. Moreover, we assessed whether participants complied with the instructions to use their left and right index fingers for the left and right response keys, respectively, and only included their datasets in our analysis if they did so. Then, as a manipulation check, we tested the SNARC effect in the lower number ranges (hypothesis 1a). Importantly, we only proceeded with the testing of other hypotheses because we could find the SNARC effect in the lower number range in both experiments.

3. Possible limitations and unexpected outcomes

Finding evidence against the SNARC effect in the lower number ranges (Experiment 1: 0–5; Experiment 2: 1–5 , excluding 3) would have been an unexpected outcome. The SNARC effect in the parity judgement task has been shown in plenty of studies (including online experiments) using different number ranges within the interval from 0 to 9. Our large sample and a high number of repetitions ensured a high probability of finding evidence even for small effects. The presence of the SNARC effect in the lower ranges thus is a manipulation check and prerequisite for further hypothesis tests.

Even though our Experiment 1 was a direct replication of Dehaene *et al.*'s [6] and Fias *et al.*'s [8] study, we decided to use 25 instead of 15 repetitions per experimental cell. First, we thereby increased statistical power and measurement precision [73]; second, we followed methodological recommendations for investigating the SNARC effect [34]; and third, we ensured the comparability with our conceptual replication in Experiment 2. Because of this methodological improvement, our experiment was not a direct replication.

Just as the original two experiments, Experiment 1 had the limitation of the MARC effect being confounded with the SNARC effect because number parity and number magnitude were no orthogonal predictors in the regression model. Therefore, we only calculated the MARC effect for the data resulting from Experiment 2. Moreover, because of the special characteristics of the number 0 regarding its mental representation, including it in the stimulus set might have driven responses in our Experiment 1. However, we tackled these limitations in Experiment 2 by using another stimulus set.

4. Results

The data collection as well as the confirmatory data analyses were conducted as described in the peer-reviewed Stage-1 Registered Report, which received in-principle acceptance by the

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

Table 2. This Study Design Table contains all research questions with corresponding hypotheses, the targeted sample size and planned analyses with a rationale, as well as the interpretations of potential outcomes and theoretical conclusions. All entries apply to both Experiment 1 (direct replication using 0–5 and 4–9) and Experiment 2 (conceptual replication using 1–5 excluding 3 and 4 to 8 excluding 6). The template for the Study Design Table was taken from PCI-RR and filled in before data collection started. Only the rightmost column, 'bserved outcome' has been added after data was collected and analyses were run.

Question	Hypothesis	Analysis plan	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes	Observed outcome
Can a SNARC effect be observed in all number ranges?	*Hypothesis 1 (and manipulation check):* A robust SNARC effect is expected in both (a) the lower and (b) the higher number ranges, i.e. we expect to find at least moderate evidence for SNARC slopes (one per participant and per number range, calculated by regressing dRTs on number magnitude) to be smaller than zero in each number range. As the SNARC effect is very robust, especially for lower ranges and possibly stronger than in higher ranges, the SNARC effect in lower ranges (hypothesis 1a) will be used as manipulation check and prerequisite for following investigations (hypotheses 1b, 2a, 2b and 3).	Four regressions of dRTs on number magnitude followed by four two-sided Bayesian one-sample *t*-tests of SNARC slopes against zero (Experiment 1: 0–5 and 4–9; Experiment 2: 1–5 (excluding 3) and 4–8 (excluding 6)	Finding moderate or even strong evidence for a SNARC slope smaller than 0 in a Bayesian *t*-test in each number range would provide evidence for a SNARC effect in both the lower (hypothesis 1a) and higher (hypothesis 1b) number ranges and be in line with results from previous studies (e.g. the two seminal studies by Dehaene *et al.* [6] and by Fias *et al*. [8]).	The SNARC effect in the parity judgement task has been shown in numerous studies using different number ranges within the interval from 0 to 9 (as in all scenarios, see electronic supplementary material, figures S1 and S2 and table S1: [37]. We therefore expect to find at least moderate evidence for it in all four number ranges. Finding at least moderate evidence against the SNARC in any of the four number ranges would be highly surprising, especially in the lower number ranges. Evidence against the SNARC effect in the higher ranges (hypothesis 1b) combined with evidence for the SNARC effect in the lower ranges (hypothesis 1a) would provide support for AM dependency of the strength of the SNARC effect (hypothesis 3).	**Experiment 1:** Strong evidence for a SNARC effect in the lower range with $BF_{10} = 6956.04$ ($d = 0.38$) and in the higher range with $BF_{10} = 2.63 \times 10^{6}$ ($d = 0.47$). **Experiment 2:** Strong evidence for a SNARC effect in the lower range with $BF_{10} = 1.61 \times 10^{21}$ ($d = 0.71$) and in the higher range with $BF_{10} = 1.38 \times 10^{12}$ ($d = 0.52$). **Summary:** Manipulation check successful in both experiments, replication of previous results, prerequisite for following investigations fulfilled.
Does the number mapping on the MNL depend on whether it is the lowest versus highest number in the current number range?	*Hypothesis 2a:* For the same critical number, a left-/right-hand advantage is expected when it is the lowest/highest number in the current number range, respectively. We hypothesize RM dependency (and possibly AM dependency as well, see hypothesis 2b) of the number mapping on the MNL.	Four two-sided paired Bayesian *t*-tests of dRTs for the same number in lower versus higher number range (i.e. for 4 and 5 in each experiment. (Note that this test will only be run in case we find at least moderate evidence for a SNARC effect in the lower number range of the respective experiment; see hypothesis 1a, which serves as a manipulation check.)	Finding moderate or even strong evidence for a different pattern for numbers that appear in both number ranges in the lower and the higher number range in a *t*-test would provide evidence for RM dependency of the SNARC effect. Finding moderate or even strong evidence against a different dRT pattern would indicate AM dependency of the number mapping on the MNL.	Evidence for RM dependency would indicate flexibility of the MNL, such that its resolution adapts to the context and that RM plays a role for SNAs. However, this does not rule out the possibility that AM plays a role as well (see below). Evidence for AM dependency would indicate that the MNL is at least not fully flexible. Full RM dependency is illustrated in scenarios 1 and 4, full AM dependency is shown in scenarios 3 and 6 and a combination of both corresponds to scenarios 2 and 5 in electronic supplementary material, figures S1 and S2.	**Experiment 1:** Strong evidence for differences in dRTs for critical numbers between the ranges with $BF_{10} = 2.57 \times 10^{4}$ ($d = 0.40$) for number 4 and $BF_{10} = 12.86$ ($d = 0.25$) for number 5. **Experiment 2:** Strong evidence for differences in dRTs for critical numbers between the ranges with $BF_{10} = 6.64 \times 10^{7}$ ($d = 0.42$) for number 4 and $BF_{10} = 64.64$ ($d = 0.24$) for number 5 **Summary:** Evidence for RM dependency and thus flexibility of the number mapping on the MNL in both experiments.
Does the mapping of numbers on the MNL depend on whether they are small versus high numbers in absolute terms?	*Hypothesis 2b:* A left-/right-hand advantage could be observed for small/large numbers in absolute terms, respectively (on top of RM dependency, see hypothesis 2a). However, we cannot derive any clear hypothesis from the literature about whether dRTs are lower for the smallest number in a higher than in a lower range (as observed by Dehaene *et al*. [6], but not by Fias *et al*. [8].	Two two-sided paired Bayesian *t*-tests of smallest-number intercept in lower versus higher number range (one test per experiment). (Note that this test will only be run in case we find at least moderate evidence for a SNARC effect in the lower number range of the respective experiment; see hypothesis 1a, which serves as a manipulation check.)	Finding moderate or even strong evidence for different smallest-number intercepts in the lower and the higher number range in a Bayesian *t*-test would indicate AM dependency of the number mapping on the MNL. Finding moderate or even strong evidence against different smallest-number intercepts would indicate RM dependency of the number mapping on the MNL.	Evidence for AM dependency would indicate that the MNL and the SNARC effect are not fully flexible and that AM plays a role for SNAs. However, this does not rule out the possibility that RM plays a role as well (see above). Evidence for RM dependency would indicate that the MNL is at least partly flexible.	**Experiment 1:** Strong evidence against a difference in smallest-number intercepts between the ranges with $BF_{10} = 0.09$ **Experiment 2:** Strong evidence for a difference in smallest-number intercepts between the ranges with $BF_{10} = 546.98$ ($d = 0.27$) **Summary:** Evidence against AM dependency of the number mapping on the MNL in Experiment 1, but

(*Continued.*)

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

Table 2. (*Continued.*)

Question	Hypothesis	Analysis plan	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes	Observed outcome
					evidence for it in Experiment 2, suggesting that the number mapping on the MNL is not fully flexible when certain stimulus sets are used.
Does the strength of the SNARC effect depend on AM in the used range?	*Hypothesis 3:* The SNARC effect is expected to be stronger in the lower than in the higher number ranges.	Two two-sided paired Bayesian *t*-tests of SNARC slopes in lower versus higher number range (one test per experiment). (Note that this test will only be run in case we find at least moderate evidence for a SNARC effect in the lower number range of the respective experiment; see hypothesis 1a, which serves as a manipulation check.)	Finding moderate or even strong evidence for a more negative SNARC slope in one of the two number ranges would indicate that the SNARC effect seems to be stronger in this number range than in the other.	Finding the SNARC effect to be stronger in the lower than in the higher number range would indicate that the spatial mental representation of small numbers is more pronounced than for large numbers (as in scenarios 4, 5 and 6 in electronic supplementary material, figure S1). If the SNARC effect does not differ between number ranges, no evidence can be provided for the strength of the SNARC effect to depend on AM (as in scenarios 1, 2 and 3).	**Experiment 1:** Strong evidence against a difference in SNARC slopes between the ranges with *BF10* = 0.09. **Experiment 2:** Strong evidence for a difference in SNARC slopes between the ranges with BF_{10} = 1271.17. **Summary:** Evidence against AM dependency of the strength of the SNARC effect in Experiment 1, but evidence for it in Experiment 2, suggesting that the spatial mental representation of small numbers is more pronounced than for large numbers when certain stimulus sets are used.

Rationale for the test sensitivity and sampling plan: The most crucial aim of the present study is to find out whether AM dependency of the strength of the SNARC effect exists (*hypothesis 3*). The minimally relevant effect size of $d = 0.15$ was chosen because it corresponds to the SNARC slope difference of 2.99 ms between number ranges (with a pooled SD of 18.34 ms) that was descriptively found but remained non-significant in the original study by Fias *et al.* [8] that we wish to replicate here. Note that due to the lacking report of SD, it is not possible to calculate Cohen's *d* for the slope difference of 9.2 ms found by Dehaene *et al*. [6]. Importantly, a smaller effect size than $d = 0.15$ would not be meaningful for the SNARC effect (hypothesis 1) or for RM dependency and AM dependency of the number mapping on the MNL (hypothesis 2) either. Similarly, the chosen maximal sample size should be large enough to find at least moderate evidence in case hypotheses 1 and 2 are false.

To reach the desired probability of 0.90 for finding moderate evidence in favour of a true underlying effect (i.e. $BF_{10} \times > 3$) with an effect size of Cohen's $d = 0.15$ in two-sided Bayesian one-sample *t*-tests or in two-sided Bayesian paired *t*-tests, 800 participants need to be tested in each experiment (for power calculations and sample size estimations, see [37]. The required sample size for finding moderate evidence against a truly absent effect (i.e. $BF_{10} < 1/3$) for $d = 0$ is only 180. By ensuring our design is sensitive to find evidence for $d = 0.15$, we will be able to detect a slope difference of the size found by Fias *et al.* [8], as predicted by hypothesis 3, and a smaller effect size would not be meaningful for hypotheses 1 and 2 either. However, we will employ the SBF + maxN approach as described by Schönbrodt & Wagenmakers [54]. More precisely, we will first recruit 200 participants and then calculate the BF_{10} for all *t*-tests after each added 20 participants. In case the BF_{10} reach a threshold of 1/3 or of 3 (i.e. moderate evidence for or against hypotheses 1, 2, and 3) before getting to the sample size of 800 participants, we will stop recruiting earlier in the respective experiment.

royalsocietypublishing.org/journal/rsos *R. Soc. Open Sci.* **12**: 241585

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

PCI on 3 December 2023 (https://doi.org/10.17605/OSF.IO/AE2C8). Additional data analyses are referred to as exploratory data analyses in the following. All R scripts for data preprocessing and analysis as well as all anonymized datasets can be found at [37]. A Study Design Table was filled in prior to data collection and provides an overview of all research questions, corresponding hypotheses, the targeted sample size and planned analyses with a rationale, the interpretations of potential outcomes and theoretical conclusions (see table 2). It also contains the observed outcomes for both experiments, which are additionally illustrated in figure 3.

4.1. Experiment 1: close replication with 0 to 5 versus 4 to 9

4.1.1. Data preprocessing

Our final recruited sample size according to the SBF + maxN approach was 200. This final sample size refers only to collected individual datasets that were complete, but initially, 208 individuals started participation of which eight did not complete. Of these, one participant was under 18 and two participants were over 40 years old. Three participants did not fully follow the instructions as concerns response-to-key assignment, and seven participants used other fingers than required in the instructions or switched around. The data of these participants were excluded from the analysis in the first preprocessing step. All recruited individuals wanted to seriously participate (i.e. none of them only wanted to look at the experiment), and none of them reported a very/extremely noisy environment or multiple major distractions. In the next preprocessing step, exclusion criteria were applied at the trial level. That is, 0.01% of the trials were excluded due to missing responses, 5.78% due to incorrect responses, 0.12% due to responses faster than 200 ms, 2.17% due to responses slower than 1500 ms and 6.14% in the sequential RT trimming procedure. After these exclusions at the trial level, none of the participants had any empty experimental cells, but 14 participants had less than 75% remaining valid trials and their data were thus excluded from analyses. The remaining data of 173 individuals were analysed. Descriptive self-reported information about these participants is summarized in table 3. All of them used the default response keys D and K. The average RT per number in each range can be found in table 4 and is illustrated in electronic supplementary material, figures S3 and S4. Mean RTs ranged from 525.25 to 567.35 ms and standard errors (SE) ranged from 6.75 to 8.43 ms (note that the descriptively largest SE was observed for number 0, which is in line with previous studies).

4.1.2. Confirmatory data analysis

All parameter estimates can be found in table 5.[2] The average RT for all included trials was 543.44 ms (SD = 88.00 ms). The repeated-measures regressions followed by Bayesian one-sample *t*-tests against zero revealed strong evidence for a SNARC effect in the lower range (i.e. for hypothesis 1a) with BF_{10} = 6956.04 and a slope estimate of −5.53 ms (SD = 14.66 ms, d = 0.38), as well as in the higher range (i.e. for hypothesis 1b) with $BF_{10} = 2.63 \times 10^6$ and a slope estimate of −5.84 ms (SD = 12.33 ms, d = 0.47). Hence, the manipulation check confirmed that the manipulation had worked, and the prerequisite for testing all further hypotheses was fulfilled. The SNARC effect is plotted in figure 4 for both ranges (see also figure 3, left panel) and looks similar to scenario 1 illustrated in electronic supplementary material, figure S1 (see electronic supplementary material in 'PCI Stage 2 Registered Report Materials' at [37]. Two paired Bayesian *t*-tests revealed strong evidence for differences in dRTs for critical numbers between the ranges and thus for RM dependency of the number mapping on the MNL (i.e. for hypothesis 2a) with $BF_{10} = 2.57 \times 10^4$ (d = 0.40) for number 4 and BF_{10} = 12.86 (d = 0.25) for number 5. Another paired Bayesian *t*-test revealed strong evidence against a difference in smallest-number intercepts between the ranges and thus against AM dependency of the number mapping on the MNL (i.e. against hypothesis 2b) with BF_{10} = 0.09. The last paired Bayesian *t*-test revealed strong evidence against a difference in SNARC slopes between the ranges and thus against AM dependency of the strength of the SNARC effect (i.e. against hypothesis 3) with BF_{10} = 0.09.

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

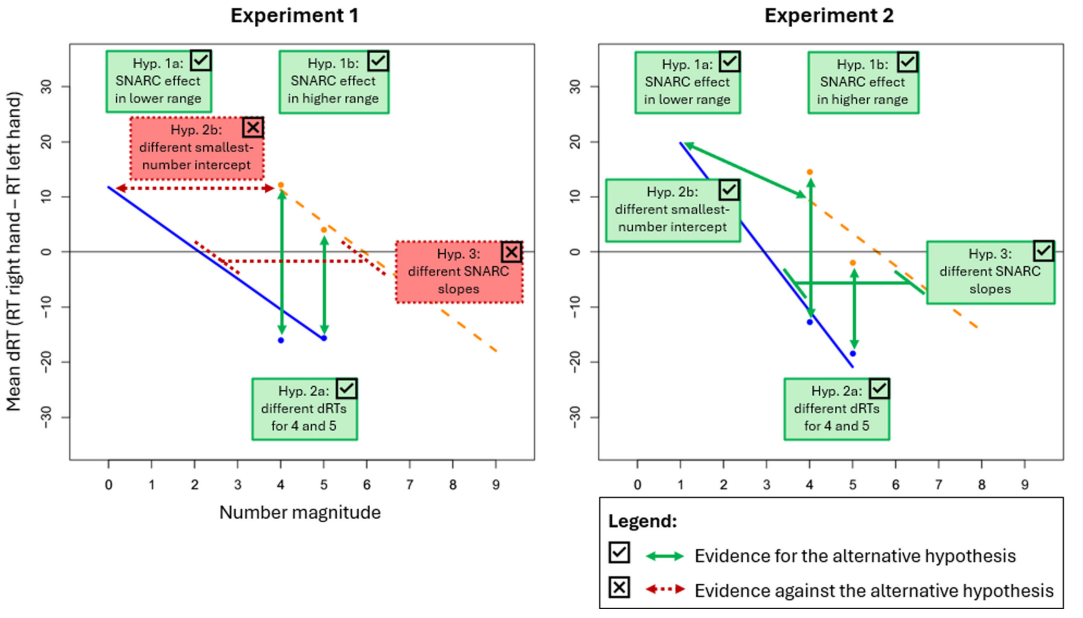


Figure 3. Summary of all tested hypotheses and outcomes in the plot for the linear regression of mean dRTs on number magnitude separately for the lower (blue, solid line) and higher (orange, dashed line) number ranges for experiments 1 (left; see figure 4) and 2 (right; see figure 5). *Note*. The figure only includes the mean dRTs for the critical numbers 4 and 5, which appear in both the lower and the higher number ranges. Hypotheses 1a and 1b were tested with one-sample *t*-tests, whereas hypotheses 2a, 2b and 3 were comparisons tested with paired *t*-tests and are illustrated with two-sided arrows. Green boxes with a solid outline and a checkmark as well as green solid arrows indicate Bayesian evidence for the alternative hypothesis (i.e. $BF_{10} > 3$). Red boxes with a dotted outline and a cross as well as red dotted arrows indicate Bayesian evidence against the alternative hypothesis (i.e. $BF_{10} < 1/3$).

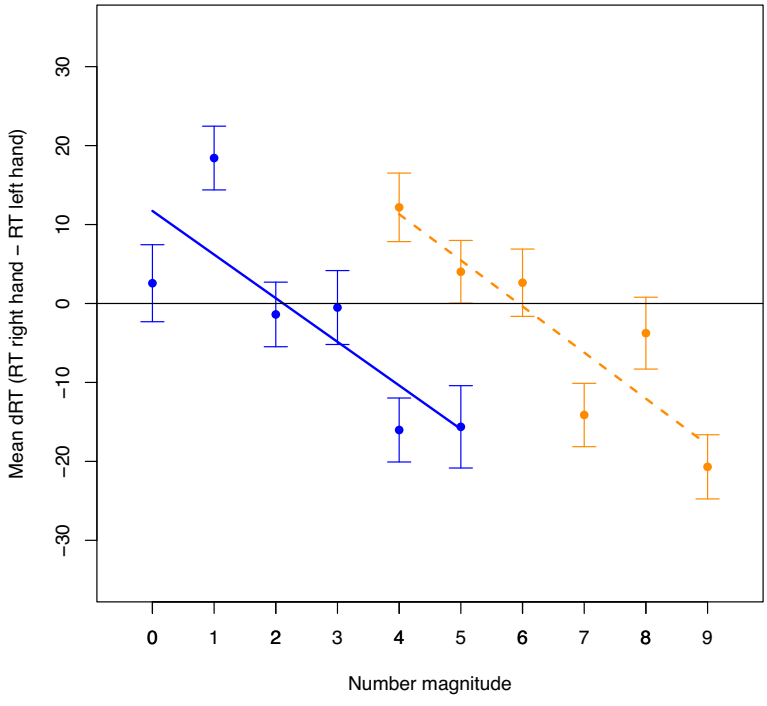


Figure 4. Mean dRTs per number averaged across all trials from all participants in the lower (blue, solid line) and higher (orange, dashed line) number ranges for Experiment 1, with error bars representing ± 1 SE for the respective number and regression lines representing slope estimates for the respective range.

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

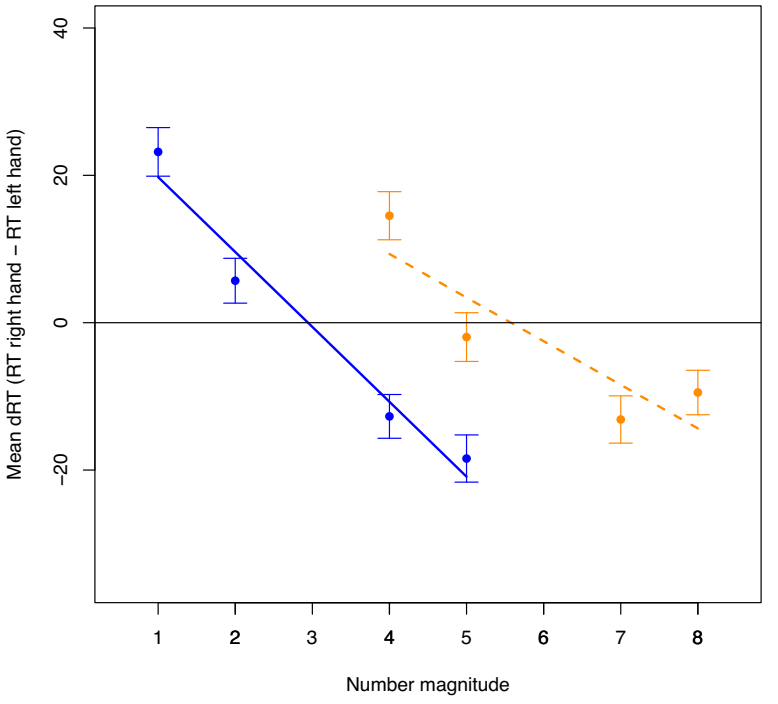


Figure 5. Mean dRTs per number averaged across all trials from all participants in the lower (blue, solid line) and higher (orange, dashed line) number ranges for Experiment 2, with error bars representing ± 1 SE for the respective number and regression lines representing slope estimates for the respective range.

Table 3. Descriptive self-reported information about the samples in both experiments ($N = 173$ in Experiment 1 and $N = 255$ in Experiment 2 after exclusions).

demographic item	answer	Experiment 1	Experiment 2
gender	woman	79 (45.7%)	121(47.5%)
	man	92 (53.2%)	133 (52.2%)
	diverse	2 (1.2%)	1 (0.4%)
	no answer	0 (0.0%)	0 (0.0%)
age	mean (SD)	30.10 (5.52)	30.20 (5.86)
native language	English	171 (98.8%)	253 (99.2%)
	others	2 (1.2%)	2 (0.8%)
handedness	right-handed	150 (86.7%)	224 (87.8%)
	left-handed	15 (8.7%)	24 (9.4%)
	ambidextrous	8 (4.6%)	7 (2.7%)
finger-counting habit	right-starters	82 (47.4%)	128 (50.2%)
	left-starters	64 (37.0%)	95 (37.3%)
	does not know or no preference	27 (15.6%)	32 (12.5%)
finger-counting stability	always	57 (32.9%)	99 (38.8%)
	mostly	72 (41.6%)	103 (40.4%)
	slightly more often	15 (8.7%)	15 (5.9%)
	does not know or no preference	29 (16.8%)	38 (14.9%)
math skills (0–400)	mean (SD)	240.80 (86.94)	235.80 (94.32)

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

Table 4. Average RT in ms per number in each range (with SEs in parentheses). For plots, see electronic supplementary material, figures S3–S6.

Experiment 1				Experiment 2			
lower range		higher range		lower range		higher range	
number	mean RT (SE RT)	number	mean RT (SE RT)	number	mean RT (SE RT)	number	mean RT (SE RT)
0	566.63 (8.43)	4	525.25 (6.75)	1	517.87 (5.19)	4	513.47 (5.15)
1	567.35 (7.58)	5	544.53 (6.97)	2	512.08 (5.26)	5	519.53 (5.07)
2	538.28 (7.61)	6	540.12 (7.40)	4	514.18 (5.24)	7	508.70 (5.12)
3	547.82 (7.59)	7	529.12 (7.00)	5	522.44 (5.31)	8	501.59 (5.12)
4	539.00 (7.11)	8	529.96 (6.92)	—	—	—	—
5	550.15 (7.52)	9	548.12 (7.28)	—	—	—	—

4.1.3. Exploratory data analysis

In addition to the confirmatory data analyses and in order to disentangle the possible scenarios illustrated in electronic supplementary material, figures S1 and S2 and table S1, the mean-number intercepts were compared between ranges. The mean number in the lower range (0–5) was 2.5 with a dRT estimate of −2.09 ms, and the mean number in the higher range (4–9) was 6.5 with a dRT estimate of −3.29 ms. A two-sided paired Bayesian *t*-test revealed moderate evidence against a difference in mean-number intercepts between the ranges and thus against AM dependency of the number mapping on the MNL with $BF_{10} = 0.10$.

Moreover, we tested whether there was a correlation between the SNARC slopes in the lower and the higher range. The data revealed moderate evidence against a correlation with $BF_{10} = 0.19$.

4.2. Experiment 2: conceptual replication with 1–5 (excluding 3) versus 4–8 (excluding 6)

4.2.1. Data preprocessing

Our final recruited sample size according to the SBF + maxN approach was 300.[3] This final sample size refers only to collected individual datasets that were complete, but initially 310 individuals started participation of which 10 did not complete. Of these, two individuals did not want to seriously participate and instead only looked at the experiment, one was over 40 years old, two participants reported a very/extremely noisy environment and three participants reported multiple major distractions. Further, 17 participants did not fully follow the instructions as concerns response-to-key assignment, and 10 participants used other fingers than required in the instructions or switched around. The data of these participants were excluded from the analysis in the first preprocessing step. No participant was under 18 and needed to be removed. In the next preprocessing step, exclusion criteria were applied at the trial level. 0.37% of the trials were excluded due to missing responses, 5.52% due to incorrect responses, 0.47% due to responses faster than 200 ms, 1.22% due to responses slower than 1500 ms and 5.49% in the sequential RT trimming procedure. After these exclusions at the trial level, none of the participants had any empty experimental cells, but 15 participants had less than 75% remaining valid trials, and their data were thus excluded from analyses. The remaining data of 255 individuals were analysed. Descriptive self-reported information about these participants is summarized in table 3. All participants used the default response keys D and K. The average RT per number in each range can be found in table 4 and is illustrated in electronic supplementary material, figures S5 and S6. Mean RTs ranged from 501.59 to 522.44 ms and SEs ranged from 5.07 to 5.31 ms.

4.2.2. Confirmatory data analysis

All parameter estimates can be found in table 5. The average RT for all included trials was 512.77 ms (SD = 74.46 ms) (table 5). The repeated-measures regressions followed by Bayesian one-sample *t*-tests against zero revealed strong evidence for a SNARC effect in the lower range (i.e. for hypothesis 1a) with $BF_{10} = 1.61 \times 10^{21}$ and a slope estimate of −10.17 ms (SD = 14.32 ms, $d = 0.71$), as well as in the higher range (i.e. for hypothesis 1b) with $BF_{10} = 1.38 \times 10^{12}$ and a slope estimate of −5.92 ms (SD = 14.32

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

Table 5. Parameter estimates for both experiments in the lower and higher number ranges (calculated per participant and averaged across them), with one asterisk indicating moderate evidence and with two asterisks indicating strong evidence for H0 (i.e. no difference between ranges) or for H1 (i.e. difference between ranges).

	Experiment 1			Experiment 2		
	lower	higher	evidence	lower	higher	evidence
SNARC intercept	11.72	34.68	H1**	29.93	33.00	H0*
SNARC slope	−5.53	−5.84	H0*	−10.17	−5.92	H1**
dRT for number 4	−16.03	12.18	H1**	−12.72	14.52	H1**
dRT for number 5	−15.63	4.02	H1*	−18.44	−1.95	H1**
smallest-number intercept	11.72	11.31	H0*	19.76	9.32	H1**
mean-number intercept	−2.09	−3.29	H0*	−0.57	−2.51	H0*
MARC slope	—	—	—	—	10.07	inconcl.

ms, $d = 0.52$). Hence, the manipulation check worked (i.e. evidence for hypothesis 1a was obtained), and the prerequisite for testing all further hypotheses was fulfilled. The SNARC effect is plotted in figure 5 for both ranges (see also figure 3, right panel) and seems to correspond to scenario 5 presented in electronic supplementary material, figure S2. Two paired Bayesian *t*-tests revealed strong evidence for differences in dRTs for critical numbers between the ranges and thus for RM dependency of the number mapping on the MNL (i.e. for hypothesis 2a) with $BF_{10} = 6.64 \times 10^7$ ($d = 0.42$) for number 4 and $BF_{10} = 64.64$ ($d = 0.24$) for number 5. Another paired Bayesian *t*-test revealed strong evidence for a difference in smallest-number intercepts between the ranges and thus for AM dependency of the number mapping on the MNL (i.e. for hypothesis 2b) with $BF_{10} = 546.98$ ($d = 0.27$). The last paired Bayesian *t*-test revealed strong evidence for a difference in SNARC slopes between the ranges and thus for AM dependency of the strength of the SNARC effect (i.e. for hypothesis 3) with $BF_{10} = 1271.17$ ($d = 0.28$).[4]

4.2.3. Exploratory data analysis

As for Experiment 1, to disentangle the possible scenarios illustrated in electronic supplementary material, figures S1 and S2 and table S1 in Experiment 2, the mean-number intercepts were compared between ranges. The mean number in the lower range (1–5, excluding 3) was 3 with a dRT estimate of −0.57 ms, and the mean number in the higher range (4–8, excluding 6) was 6 with a dRT estimate of −2.51 ms. A two-sided paired Bayesian *t*-test revealed moderate evidence against a difference in mean-number intercepts between the ranges and thus against AM dependency of the number mapping on the MNL with $BF_{10} = 0.18$.

As in Experiment 1, we tested whether there was a correlation between the SNARC slopes in the lower and the higher range. In contrast to the data of Experiment 1, the data of Experiment 2 revealed strong evidence for a moderate correlation with an estimate of $r = 0.34$ and $BF_{10} = 5.22 \times 10^5$.

In contrast to Experiment 1, number parity and number magnitude were orthogonal in Experiment 2 (i.e. the mean number magnitude was equal for odd and even numbers in each range). Therefore, we were also able to test the MARC effect in Experiment 2. A two-sided Bayesian one-sample *t*-test of the MARC slopes against zero revealed moderate evidence against a MARC effect in the lower range with $BF_{10} = 0.16$ and inconclusive evidence regarding a MARC effect in the higher range with $BF_{10} = 0.51$.

5. Discussion

The goal of the present study was to determine the degree of flexibility of SNAs. More precisely, we wanted to find out whether the SNARC effect is entirely flexible and depends only on relative magnitude (RM dependency) or whether it is less flexible than previously assumed and also depends on absolute magnitude (AM dependency). Importantly, the concepts of RM dependency versus AM dependency can be differentiated in the following two ways: (i) the number mapping on the MNL

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

(e.g. dRT for number 4) can be RM-dependent, AM-dependent or both, and (ii) the strength of the SNARC effect in terms of the relative increase of right-hand advantage per increase in magnitude (i.e. the SNARC slope) can be AM dependent or not. To summarize, the aim of the study was to determine whether SNAs operate on fixed and flexible number representations simultaneously.

5.1. RM dependency and AM dependency of SNAs

In the two seminal studies by Dehaene *et al*. [6] (Experiment 3) and by Fias *et al*. [8] (Experiment 1), numbers 4 and 5 were associated with the right when presented in the range from 0 to 5, but with the left when presented in the range from 4 to 9. These results are often quoted in the literature as evidence for pure RM dependency, although the studies were underpowered for small effects and descriptive data suggests AM dependency as well. In Experiment 1, which was run online with the same stimulus sets as in the two original studies, we replicated the findings from the original studies. Specifically, we found strong evidence for a small-sized SNARC effect in the lower (average slope of −5.53) and higher (average slope of −5.84) range separately (hypothesis 1), as well as for small dRT differences for critical numbers (i.e. 4 and 5) between the ranges (hypothesis 2a), in line with the original studies. Moreover, we observed strong evidence against smallest-number intercept (hypothesis 2b) and SNARC slope differences (hypothesis 3), as well as moderate evidence against mean-number intercept differences (exploratory analysis) between the lower and the higher range. To conclude, the results from our investigation are entirely in line with the results from the original studies, and in fact, without further consideration (see below), they seem to support the conclusion of full RM dependency drawn by Dehaene *et al*. [6] and Fias *et al*. [8]: whereas their data hinted descriptively towards AM dependency as pointed out in the Introduction, we even observed moderate evidence against it.

Crucially, the second experiment in the present study yielded different results. In Experiment 2, we ran a conceptual online replication considering recent advances in SNARC research with the number ranges from 1 to 5 (excluding 3) and from 4 to 8 (excluding 6). As opposed to the original stimulus set, this modified stimulus set avoids potential confounds with the MARC effect thanks to the equal number of odd and even numbers in the original stimulus set and parity being orthogonal to magnitude. Also, number 0 was excluded because it has special properties and deviates in its dRT pattern from all other numbers (see [41] for detailed analysis). In line with the literature, the dRT pattern for number 0 observed in the current study did not align well with the regression line, and the RT variation was descriptively larger for number 0 than for all other numbers. Therefore, one would not want to base a general statement about the flexibility of the number line on number 0, because the relatively low dRT estimate for 0 might considerably attenuate the SNARC effect in the lower range. However, these two number ranges used in Experiment 2 still include the same critical numbers that are part of both number ranges in the original studies, namely 4 and 5. As in Experiment 1, we found strong evidence for the SNARC effect in the lower (average slope of −10.17) and higher (average slope of −5.92) range separately (hypothesis 1). Note that in contrast to the small effect size in both ranges in Experiment 1 ($d = 0.38$ and $d = 0.47$), the effect size was medium in both ranges in Experiment 2 ($d = 0.71$ and $d = 0.52$). Thus, our results support the claim that the inclusion of number 0 in the stimulus set or a potential confound with the MARC effect due to an unequal number of odd and even numbers which might have decreased the SNARC effect in the seminal studies by Dehaene *et al*. [6] and Fias *et al*. [8]. Moreover, as in Experiment 1, strong evidence was found for dRT differences for critical numbers (i.e. 4 and 5) between ranges (hypothesis 2a). Further, as in Experiment 1, moderate evidence was found against mean-number intercept differences between ranges (exploratory analysis). However, the support for RM dependency is not the entire story.

Importantly, the data revealed strong evidence for differences in smallest-number intercept (hypothesis 2b) and SNARC slopes (hypothesis 3) between the ranges. These differences indicating AM dependency are only small ($d = 0.27$ and $d = 0.28$ for hypotheses 2b and 3, respectively), but so are the differences indicating RM dependency ($d = 0.42$ and $d = 0.24$ for hypothesis 2a). Hence, in contrast to the pure RM dependency observed in Experiment 1, the results from Experiment 2 suggest AM dependency both for (i) the number mapping on the MNL and (ii) the strength of the SNARC effect (in line with scenario 5 in the electronic supplementary material).

How can these results be reconciled with Experiment 1, where we observed evidence against AM dependency? As outlined above, we conducted Experiment 2 (among other reasons) to see whether 0 drives results for the number ranges from 0 to 5 and from 4 to 9 used in the original experiments by Dehaene *et al*. [6] and Fias *et al*. [8]. Therefore, we reanalysed the data in an exploratory way without 0. Note this is not the same as if the experiment had been run without number 0. It was part of the

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

stimulus set, even when it was not analysed later, and the number range used in the experiment might still influence results. When excluding number 0 from the stimuli set in Experiment 1 post hoc,[1] the evidence against AM dependency disappeared. Instead, the evidence was now inconclusive regarding smallest-number intercept differences (hypothesis 2b) and SNARC slope differences (hypothesis 3). Since number 0 was still part of the range of Experiment 1, we did not expect the same results as in Experiment 2, which was run without 0. However, when the results of Experiment 1 are analysed without 0, there is at least no conflicting evidence anymore. This change suggests that the inclusion of 0 in some range plays a major role for the eventual outcome.

Notably, the dRT was also rather high for number 1 in both experiments and pulled the regression line upwards. That is, number 1 seems to be more strongly associated with the left than what would be predicted based on the regression slope alone. This observation fits with the results from our two previously conducted colour judgement tasks, which do not require semantic number processing at all [32]. Number 1 seemed to be strongly associated with the left in these experiments as well, providing further support for AM dependency of the MNL.

In the present study, Bayesian analyses allowed us to interpret and quantify evidence for the null hypotheses. Moreover, these findings can be considered as trustworthy, because the sample sizes were large enough (173 datasets analysed out of 200 recruited participants in Experiment 1 and 255 datasets analysed out of 300 recruited participants in Experiment 2). Thanks to the SBF + maxN approach, an optimally sized sample was recruited in each experiment. These samples were much larger than in the original studies, where a difference between number ranges might just have stayed undetected due to lacking statistical power.

To sum up, although the picture is blurred by methodological issues, especially the inclusion of number 0 in one but not in the other range, in the original studies and in our Experiment 1, the findings from Experiment 2 together with the reanalysis of Experiment 1 without 0 seem to suggest that there is not the 'one and only SNARC'. The spatial mapping of numerical magnitude onto space seems not to be fully flexible and dependent on the used range. Null effects of AM can only be found when number 0 is included. We therefore conclude that the SNARC effect seems to operate on multiple number representations and on multiple spatial reference frames simultaneously, namely on both flexible and absolute ones.

5.2. RM dependency and AM dependency from a theoretical point of view

As outlined in the Introduction, different predictions regarding the SNARC effect's flexibility can be derived from the models that have been proposed to account for the origin of the SNARC effect. The working memory account [28,29] postulates that the SNARC effect is constructed during task execution. In the present study, the currently used stimulus set (i.e. the lower or the higher number range) was stored in working memory, which was reflected by the differential patterns for critical numbers 4 and 5. Namely, when these critical numbers were the highest (i.e. in the lower ranges in both experiments), they were associated with the right, whereas they were associated with the left when they were the lowest (i.e. in the higher ranges in both experiments). This observed RM dependency is clearly in line with the working memory account. In contrast, although the MNL [6] is claimed to dynamically adapt to task demands as well, such that zooming in and out is possible [17], it can be considered as a mental representation in long-term memory. Similarly, the verbal–spatial coding account [25] and polarity correspondence account [26] postulate number representations stored in long-term memory. Crucially, long-term representations hardly justify the SNARC effect's flexibility [15,27]. Instead, they are in line with the AM dependency observed in Experiment 2, reflected by (i) the number mapping on the MNL in terms of smallest-number intercept differences between ranges and (ii) the degree of spatialization in terms of differences between ranges regarding the strength of the SNARC effect. A potential explanation for this AM dependency is that small numbers are more frequently used than large numbers, which might lead to a more fixed and stronger spatial mental representation. Importantly, RM dependency and AM dependency can coexist (in line with [14,27,30]) and multiple spatial reference frames can be activated simultaneously [31].

An MNL account, in which SNAs are partly flexible (aka zoomed in) and partly fixed (i.e. always more left for absolutely smaller numbers), can also explain the current data (see [30]). However, any account, which wishes to explain the current data, needs a fixed and a flexible component. Any fully flexible account is in our view not consistent with these data.

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

5.3. Correlation of the SNARC effect between ranges

Apart from the SNARC effect at the group level, one can also investigate the phenomenon at the individual level [47,74]. Assessing parity judgement in both stimulus ranges in a within-subjects design permits to investigate the correlation of SNARC effects between ranges (exploratory analyses), which was not tested in the original studies. Surprisingly, Experiment 1 revealed moderate evidence against a correlation of SNARC slopes between ranges, whereas Experiment 2 revealed strong evidence for a moderate correlation, with an estimate of $r = 0.34$. A hypothetical explanation could be that the inclusion of number 0 in the stimulus set of Experiment 1 might have blurred a true underlying correlation.[5] That is, number 0 showed descriptively stronger variations in RTs than other numbers did, which is in line with the literature (see figure 4 in Nuerk *et al*. [41]), thus introducing a large error term for the dRT. The dRT for number 0 is rather uncorrelated with dRTs for other numbers [41], so slopes built on a stimulus set including 0 are also more likely to be uncorrelated with other slopes built on stimulus sets excluding 0. This large variation in turn results in increased noise for the SNARC slopes in the lower number range in Experiment 1. To summarize, our results from Experiment 2 show a positive relationship between the strength of the spatial mapping in small and large numbers within participants.

6. Conclusion

The current study demonstrates that the spatial mental representation of numbers is not entirely flexible. Mental spatialization can be adapted to the context and depends largely on RM, but at the same time, it is also influenced by AM. RM and AM play a role in (i) the association of specific numbers with horizontal directional space, such that numbers that are small in relative or absolute terms are associated with the left, whereas numbers that are large in relative terms are associated with the right. At the same time, RM and AM play a role in (ii) the strength of the spatialization, such that the association of small with left and large with right is stronger within a lower than within a higher number range. To conclude, the spatial representation of number magnitude seems to be partly flexible and partly fixed.

Ethics. This study has been approved by the ethics committee of the University of Tübingen's Department of Psychology. The Stage-1 Registered Report received in-principle acceptance by the Peer Community on 3 December 2023, and the Stage-2 Registered Report received the final recommendation on 6 September 2024 (https://doi.org/10.24072/pci.rr.100808).

Data accessibility. Demo version for Experiment 1 is available at https://exp.wextor.eu/esnarc/flex/?demo&e1 and for Experiment 2 at https://exp.wextor.eu/esnarc/flex/?demo&e2).

All R scripts for data preprocessing and data analysis, as well as all anonymized datasets are available online at [37].

Supplementary material is available online [75].

Declaration of AI use. The authors declare they have not used AI-assisted technologies in creating this article.

Authors' contributions. L.R.: data curation, formal analysis, investigation, methodology, project administration, validation, visualization, writing—original draft; J.C.: software; U.-D.R.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, writing—review and editing; H.-C.N.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing—review and editing; A.T.O.: formal analysis, investigation, software, writing—review and editing; K.C.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, supervision, validation, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. The authors declare they have no competing interests.

Funding. This research was supported by the DFG project 'Replicability of Fundamental Results on Spatial-Numerical Associations in Highly Powered Online Experiments (e-SNARC)' (NU 265/8-1 and RE 2655/3-1) granted to Hans-Christoph Nuerk and Ulf-Dietrich Reips supporting Lilly Roth, John Caffier, and Annika Tave Overlander, with the assistance of Krzysztof Cipora as a cooperation partner. Krzysztof Cipora is supported by the UKRI Economic and Social Research Council (grant number ES/W002914/1).

Acknowledgements. The authors would like to thank Sebastian Sandbrink for English proofreading of the stage-1 Registered Report. Moreover, the authors wish to thank PCI-RR and especially the recommender Robert McIntosh for their helpful feedback at both stages 1 and 2.

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

Endnotes

[1]Note that Prolific claims that only one participation from each household is allowed per study. Nevertheless, we received two different complete individual datasets with the same User-Agent, IP address and Prolific ID. The analyses reported in the current paper are based on both datasets. However, we reran the analyses while excluding the second of these two individual datasets and also both of these individual datasets. Importantly, apart from small changes in the decimals of the obtained Bayes factors, the results and interpretations did not change substantially in either of the two reanalyses.

[2]When excluding number 0 from the lower and number 4 from the higher range in the analyses, the results of the confirmatory data analysis did not change substantially for hypothesis 1a with $BF_{10} = 2.09 \times 10$ and for hypothesis 1b with $BF_{10} = 6.65 \times 10$, and naturally remained the same for the dRT of number 5 in hypothesis 2a. However, for hypotheses 2b and 3, the exclusion of number 0 from the lower and number 4 from the higher range led to inconclusive evidence regarding AM dependency of the number mapping on the MNL and of the strength of the SNARC effect, with $BF_{10} = 2.56$ and $BF_{10} = 0.50$, respectively. This stands in contrast with moderate evidence against AM dependency found when including number 0 in the lower and number 4 in the higher range.

[3]Note that due to a computing mistake, data collection was only stopped with 300 participants, although the stopping criterion for the SBF + maxN procedure (i.e. at least moderate Bayesian evidence for or against each hypothesis) was already reached with 200 participants. Importantly, using data from 300 instead of only 200 participants did not change any results substantially and mostly made the Bayesian evidence stronger (see next footnote).

[4]If data collection had been stopped with a sample size of 200 participants, results would have been similar with $BF_{10} = 4.13 \times 10$ ($d = 0.70$) for hypothesis 1a, $BF_{10} = 2.34 \times 10$ ($d = 0.57$) for hypothesis 1b, $BF_{10} = 5563.96$ ($d = 0.37$) for number 4 and $BF_{10} = 75.13$ ($d = 0.29$) for number 5 for hypothesis 2a, $BF_{10} = 33.56$ ($d = 0.27$) for hypothesis 2b and $BF_{10} = 19.46$ ($d = 0.26$) for hypothesis 3. Thus, evidence would also have been strong and in favour of each hypothesis with 200 participants, just like with 300 participants.

[5]Excluding number 0 from the lower and number 4 from the higher range in Experiment 1 in the analysis did not affect the results regarding the correlation: moderate evidence against a correlation was found, with $BF_{10} = 0.22$. Note that this does not contradict the strong evidence for a correlation in Experiment 2, because number 0 was part of the stimulus set when running Experiment 1 and might have undermined the potential correlation.

References

1. Cipora K, He Y, Nuerk HC. 2020 The Spatial-Numerical Association of Response Codes effect and math skills: Why related? *Ann. NY Acad. Sci.* **1477**, 5–19. (doi:10.1111/nyas.14355)
2. Cipora K, Schroeder PA, Soltanlou M, Nuerk HC. 2018 More space, better mathematics: Is space a powerful tool or a cornerstone for understanding arithmetic? In *Visualizing mathematics. Research in mathematics education* (eds KS Mix, MT Battista), pp. 77–116. Cham: Springer. (doi:10.1007/978-3-319-98767-5_4)
3. Patro K, Nuerk HC, Cress U, Haman M. 2014 How number-space relationships are assessed before formal schooling: A taxonomy proposal. *Front. Psychol.* **5**, 419. (doi:10.3389/fpsyg.2014.00419)
4. Fischer MH, Shaki S. 2014 Spatial associations in numerical cognition—from single digits to arithmetic. *Q. J. Exp. Psychol.* **67**, 1461–1483. (doi:10.1080/17470218.2014.927515)
5. Toomarian EY, Hubbard EM. 2018 On the genesis of spatial-numerical associations: Evolutionary and cultural factors co-construct the mental number line. *Neurosci. Biobehav. Rev.* **90**, 184–199. (doi:10.1016/j.neubiorev.2018.04.010)
6. Dehaene S, Bossini S, Giraux P. 1993 The mental representation of parity and number magnitude. *J. Exp. Psychol.* **122**, 371–396. (doi:10.1037/0096-3445.122.3.371)
7. Cipora K, Soltanlou M, Reips UD, Nuerk HC. 2019 The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behav. Res. Methods* **51**, 1676–1692. (doi:10.3758/s13428-019-01213-5)
8. Fias W, Brysbaert M, Geypens F, d'Ydewalle G. 1996 The importance of magnitude information in numerical processing: Evidence from the SNARC effect. *Math. Cogn.* **2**, 95–110. (doi:10.1080/135467996387552)
9. Wood G, Willmes K, Nuerk HC, Fischer MH. 2008 On the cognitive link between space and number: A meta-analysis of the SNARC effect. *Psychol. Sci.* **50**, 489–525.
10. Lorch RF, Myers JL. 1990 Regression analyses of repeated measures data in cognitive research. *J. Exp. Psychol.* **16**, 149–157. (doi:10.1037//0278-7393.16.1.149)
11. Cipora K, Patro K, Nuerk HC. 2018 Situated influences on Spatial–Numerical Associations. In *Spatial biases in perception and cognition* (ed. T Hubbard), pp. 41–59. Cambridge: Cambridge University Press. (doi:10.1017/9781316651247)
12. Antoine S, Gevers W. 2016 Beyond left and right: Automaticity and flexibility of number-space associations. *Psychon. Bull. Rev.* **23**, 148–155. (doi:10.3758/s13423-015-0856-x)
13. Deng Z, Chen Y, Zhu X, Li Y. 2017 The effect of working memory load on the SNARC effect: Maybe tasks have a word to say. *Mem. Cognit.* **45**, 428–441. (doi:10.3758/s13421-016-0676-x)
14. Ginsburg V, van Dijck JP, Previtali P, Fias W, Gevers W. 2014 The impact of verbal working memory on number-space associations. *J. Exp. Psychol. Learn. Mem. Cogn.* **40**, 976–986. (doi:10.1037/a0036378)

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

15. Ginsburg V, Gevers W. 2015 Spatial coding of ordinal information in short- and long-term memory. *Front. Hum. Neurosci*. **9**, 8. (doi:10.3389/fnhum.2015.00008)
16. Schwarz W, Keus IM. 2004 Moving the eyes along the mental number line: Comparing SNARC effects with saccadic and manual responses. *Percept. Psychophys*. **66**, 651–664. (doi:10.3758/BF03194909)
17. Pinhas M, Pothos EM, Tzelgov J. 2013 Zooming in and out from the mental number line: Evidence for a number range effect. *J. Exp. Psychol. Learn. Mem. Cogn*. **39**, 972–976. (doi:10.1037/a0029527)
18. Tlauka M. 2002 The processing of numbers in choice-reaction tasks. *Aust. J. Psychol*. **54**, 94–98. (doi:10.1080/00049530210001706553)
19. Ben Nathan M, Shaki S, Salti M, Algom D. 2009 Numbers and space: Associations and dissociations. *Psychon. Bull. Rev.* **16**, 578–582. (doi:10.3758/PBR.16.3.578)
20. Wühr P, Richter M. 2022 Relative, not absolute, stimulus size is responsible for a correspondence effect between physical stimulus size and left/right responses. *Atten. Percept. Psychophys*. **84**, 1342–1358. (doi:10.3758/s13414-022-02490-7)
21. Rugani R, Vallortigara G, Priftis K, Regolin L. 2015 Animal cognition. Number-space mapping in the newborn chick resembles humans' mental number line. *Science* **347**, 534–536. (doi:10.1126/science.aaa1379)
22. Di Giorgio E, Lunghi M, Rugani R, Regolin L, Dalla Barba B, Vallortigara G, Simion F. 2019 A mental number line in human newborns. *Dev. Sci*. **22**, e12801. (doi:10.1111/desc.12801)
23. Giurfa M, Marcout C, Hilpert P, Thevenot C, Rugani R. 2022 An insect brain organizes numbers on a left-to-right mental number line. *Proc. Natl Acad. Sci. USA* **119**, e2203584119. (doi:10.1073/pnas.2203584119)
24. Restle F. 1970 Speed of adding and comparing numbers. *J. Exp. Psychol*. **83**, 274–278. (doi:10.1037/h0028573)
25. Gevers W, Santens S, Dhooge E, Chen Q, Van den Bossche L, Fias W, Verguts T. 2010 Verbal-spatial and visuospatial coding of number-space interactions. *J. Exp. Psychol. Gen*. **139**, 180–190. (doi:10.1037/a0017688)
26. Proctor RW, Cho YS. 2006 Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychol. Bull*. **132**, 416–442. (doi:10.1037/0033-2909.132.3.416)
27. van Dijck JP, Ginsburg V, Girelli L, Gevers W. 2015 Linking numbers to space: From the mental number line towards a hybrid account. In *The Oxford handbook of numerical cognition* (eds RC Kadosh, A Dowker), pp. 89–105. Oxford University Press. (doi:10.1093/oxfordhb/9780199642342.013.020)
28. Fias W, van Dijck JP. 2016 The temporary nature of number-space interactions. *Can. J. Exp. Psychol*. **70**, 33–40. (doi:10.1037/cep0000071)
29. van Dijck JP, Fias W. 2011 A working memory account for spatial-numerical associations. *Cognition* **119**, 114–119. (doi:10.1016/j.cognition.2010.12.013)
30. Koch NN, Huber JF, Lohmann J, Cipora K, Butz MV, Nuerk HC. 2023 Mental number representations are spatially mapped both by their magnitudes and ordinal positions. *Collabra Psychol.* **9**, 67908. (doi:10.1525/collabra.67908)
31. Weis T, Nuerk HC, Lachmann T. 2018 Attention allows the SNARC effect to operate on multiple number lines. *Sci. Rep*. **8**, 13778. (doi:10.1038/s41598-018-32174-y)
32. Roth L, Caffier JP, Reips UD, Cipora K, Braun L, Nuerk HC. 2024 True colors SNARC: Semantic number processing is highly automatic. (doi:10.31234/osf.io/aeyn8)
33. Tzelgov J, Zohar-Shai B, Nuerk HC. 2013 On defining quantifying and measuring the SNARC effect. *Front. Psychol*. **4**, 302. (doi:10.3389/fpsyg.2013.00302)
34. Cipora K, Wood G. 2017 Finding the SNARC instead of hunting it: A 20*20 Monte Carlo investigation. *Front. Psychol*. **8**, 1194. (doi:10.3389/fpsyg.2017.01194)
35. Giner-Sorolla R, Aberson CL, Bostyn DH, Carpenter T, Conrique BG, Lewis NA, Soderberg C. 2019 Power to detect what? Considerations for planning and evaluating sample size. Open Science Framework. See https://osf.io/jnmya/.
36. Wickelmaier F. 2022 *Simulating the power of statistical tests: A collection of R examples*. ArXiv. See https://doi.org/10.48550/arXiv.2110.09836.
37. Roth L, Caffier JP, Reips UD, Nuerk HC, Overlander AT, Cipora K. 2024 SNARC flexibility: Range (in)dependency. OSF project. (doi:10.17605/OSF.IO/Z43PM)
38. Allaire J *et al*. 2022 Rmarkdown dynamic documents for R (version 2.18). R package. GitHub. https://github.com/rstudio/rmarkdown
39. Xie Y. 2022 knitr: A General-Purpose Package for Dynamic Report Generation in R (version 1.41). R package. See https://yihui.org/knitr/.
40. Morey RD, Rouder JN, Jamil T, Urbanek S, Forner K, Ly A. 2024 BayesFactor: Computation of Bayes factors for common designs (version 0.9.12-4.7). R package. See https://CRAN.R-project.org/package=BayesFactor.
41. Nuerk HC, Iversen W, Willmes K. 2004 Notational modulation of the SNARC and the MARC (Linguistic Markedness of Response Codes) effect. *Quart. J. Exp. Psychol. Sect. A* **57**, 835–863. (doi:10.1080/02724980343000512)
42. Zohar-Shai B, Tzelgov J, Karni A, Rubinsten O. 2017 It does exist! A left-to-right Spatial-Numerical Association of Response Codes (SNARC) effect among native hebrew speakers. *J. Exp. Psychol. Hum. Percept. Perform*. **43**, 719–728. (doi:10.1037/xhp0000336)
43. Brysbaert M. 1995 Arabic number reading: On the nature of the numerical scale and the origin of phonological recoding. *J. Exp. Psychol*. **124**, 434–452. (doi:10.1037/0096-3445.124.4.434)
44. Nieder A. 2016 Representing something out of nothing: The dawning of zero. *Trends Cogn. Sci.* **20**, 830–842. (doi:10.1016/j.tics.2016.08.008)
45. Pinhas M, Tzelgov J. 2012 Expanding on the mental number line: Zero is perceived as the "smallest." *J. Exp. Psychol. Learn. Mem. Cogn*. **38**, 1187–1205. (doi:10.1037/a0027390)
46. Levenson E, Tsamir P, Tirosh D. 2007 Neither even nor odd: Sixth grade students' dilemmas regarding the parity of zero. *J. Math. Behav.* **26**, 83–95. (doi:10.1016/j.jmathb.2007.05.004)

Downloaded from https://royalsocietypublishing.org/ on 08 January 2025

47. Cipora K, Dijck JP, Georges C, Masson N, Goebel SM, Willmes K, Pesenti M, Schiltz C, Nuerk HC. 2019 A Minority pulls the sample mean: On the individual prevalence of robust group-level cognitive phenomena – the instance of the SNARC effect. PsyArXiv. (doi:10.31234/osf.io/bwyr3)
48. Cleland AA, Bull R. 2019 Automaticity of access to numerical magnitude and its spatial associations: The role of task and number representation. *J. Exp. Psychol. Learn. Mem. Cogn*. **45**, 333–348. (doi:10.1037/xlm0000590)
49. Gökaydin D, Brugger P, Loetscher T. 2018 Sequential effects in SNARC. *Sci. Rep*. **8**, 10996. (doi:10.1038/s41598-018-29337-2)
50. Dienes Z. 2021 How to use and report Bayesian hypothesis tests. *Psychol. Conscious. Theory Res*. **8**, 9–26. (doi:10.1037/cns0000258)
51. Reips UD. 2000 The web experiment method: Advantages, disadvantages, and solutions. In *Psychological experiments on the Internet* (ed. MH Birnbaum), pp. 89–117. San Diego, USA: Academic Press. (doi:10.1016/B978-012099980-4/50005-8)
52. Reips UD. 2002 Standards for Internet-based experimenting. *Exp. Psychol*. **49**, 243–256. (doi:10.1026//1618-3169.49.4.243)
53. Kelter R. 2021 Analysis of type i and ii error rates of bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Comput. Stat*. **36**, 1263–1288. (doi:10.1007/s00180-020-01034-7)
54. Schönbrodt FD, Wagenmakers EJ. 2018 Bayes factor design analysis: Planning for compelling evidence. *Psychon. Bull. Rev*. **25**, 128–142. (doi:10.3758/s13423-017-1230-y)
55. Reips UD, Neuhaus C. 2002 WEXTOR: A web-based tool for generating and visualizing experimental designs and procedures. *Behav. Res. Methods. Instrum. Comput*. **34**, 234–240. (doi:10.3758/BF03195449)
56. Reips UD. 2009 Internet experiments: Methods, guidelines, metadata. human vision and electronic imaging XIV. *Proc. SPIE* **7240**, 724008. (doi:10.1117/12.823416)
57. Aust F, Diedenhofen B, Ullrich S, Musch J. 2013 Seriousness checks are useful to improve data validity in online research. *Behav. Res. Methods*. **45**, 527–535. (doi:10.3758/s13428-012-0265-2)
58. Reips UD. 2021 Web-based research in psychology. *Z. Psychol*. **229**, 198–213. (doi:10.1027/2151-2604/a000475)
59. Jenadeleh M, Zagermann J, Reiterer H, Reips UD, Hamzaoui R, Saupe D. 2023 Relaxed forced choice improves performance of visual quality assessment methods. In *Proc. 15th Int. Conf. on Quality of Multimedia Experience (QoMEX)*, Ghent. (doi:10.48550/arXiv.2305.00220)
60. Stieger S, Reips UD, Voracek M. 2007 Forced-response in online surveys: Bias from reactance and an increase in sex-specific dropout. *J. Am. Soc. Inf. Sci. Technol*. **58**, 1653–1660. (doi:10.1002/asi.20651)
61. Funke F, Reips UD. 2012 Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods* **24**, 310–327. (doi:10.1177/1525822X12444061)
62. Reips UD. 2007 The methodology of Internet-based experiments. In *The Oxford handbook of Internet psychology* (eds A Joinson, K McKenna, T Postmes, UD Reips), pp. 373–390. Oxford: Oxford University Press.
63. Birnbaum MH. 2004 Human research and data collection via the internet. *Annu. Rev. Psychol*. **55**, 803–832. (doi:10.1146/annurev.psych.55.090902.141601)
64. R Core Team. 2024 R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. See https://www.R-project.org/.
65. Barrett T. 2024 data.table: Extension of „data.frame" (version 1.15.2). R package. See https://CRAN.R-project.org/package=data.table.
66. Wickham H, François R, Henry L, Müller K, Vaughan D. 2023 dplyr: A grammar of data manipulation [R package, version 1.1.4]. See https://CRAN.R-project.org/package=dplyr.
67. Schaefer J, Opgen-Rhein R, Strimmer K. 2021 GeneNet: Modeling and inferring gene networks (version 1.2.16). R package. See https://CRAN.R-project.org/package=GeneNet.
68. Wickham H *et al*. 2024 ggplot2: Create elegant data visualisations using the grammar of graphics [R package, version 3.5.0]. See https://cran.r-project.org/package=ggplot2.
69. Lukács G, Kleinberg B, Doorn J van. 2022 neatStats: Neat and Painless Statistical Reporting (version 1.13.3). R package. See https://cran.r-project.org/web/packages/neatStats/index.html.
70. Wickham H. 2023 plyr: Tools for splitting, applying and combining [R package, version 1.8.9]. See https://CRAN.R-project.org/package=plyr.
71. Wickham H, Vaughan D, Girlich M, Ushey K. 2023 tidyr: Tidy messy data [R package, version 1.3.1]. See https://CRAN.R-project.org/package=tidyr.
72. Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ. 2011 Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspect. Psychol. Sci*. **6**, 291–298. (doi:10.1177/1745691611406923)
73. Luck S. 2019 Why experimentalists should ignore reliability and focus on precision. Luck Lab. See https://lucklab.ucdavis.edu/blog/2019/2/19/reliability-and-precision.
74. Roth L, Jordan V, Schwarz S, Willmes K, Nuerk HC, van Dijck JP, Cipora K. 2024 Don't SNARC me now! Intraindividual variability of cognitive phenomena - Insights from the Ironman paradigm. *Cognition* **248**, 105781. (doi:10.1016/j.cognition.2024.105781)
75. Roth L, Caffier J, Reips UD, Nuerk HC, Overlander AT, Cipora K. 2024 Supplementary material from: One and only SNARC? Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude. FigShare. (doi:10.6084/m9.figshare.c.7560265)

A5.2. Supplementary materials for Study 5

These supplementary materials belong to the following publication (see Appendix A5.1.):

Roth, L., Caffier, J., Reips, U.-D., Nuerk, H.-C., Overlander, A. T., & Cipora, K. (2025). One and only SNARC? Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude. *Royal Society Open Science, 12*(1), Article 241585. https://doi.org/10.1098/rsos.241585

One and only SNARC?

Spatial-Numerical Associations are not fully flexible and depend on both relative and absolute number magnitude

Supplementary Material

Lilly Roth[1], John Caffier[2], Ulf-Dietrich Reips[2], Hans-Christoph Nuerk[1,4,5], Annika Tave Overlander[2], Krzysztof Cipora[3*]

[1]Department of Psychology, University of Tübingen, Germany

[2]Department of Psychology, University of Konstanz, Germany

[3]Centre for Mathematical Cognition, Loughborough University, United Kingdom

[4]LEAD Graduate School & Research Network, University of Tübingen, Germany

[5]German Center for Mental Health (DZPG)

*Corresponding author:

k.cipora@lboro.ac.uk

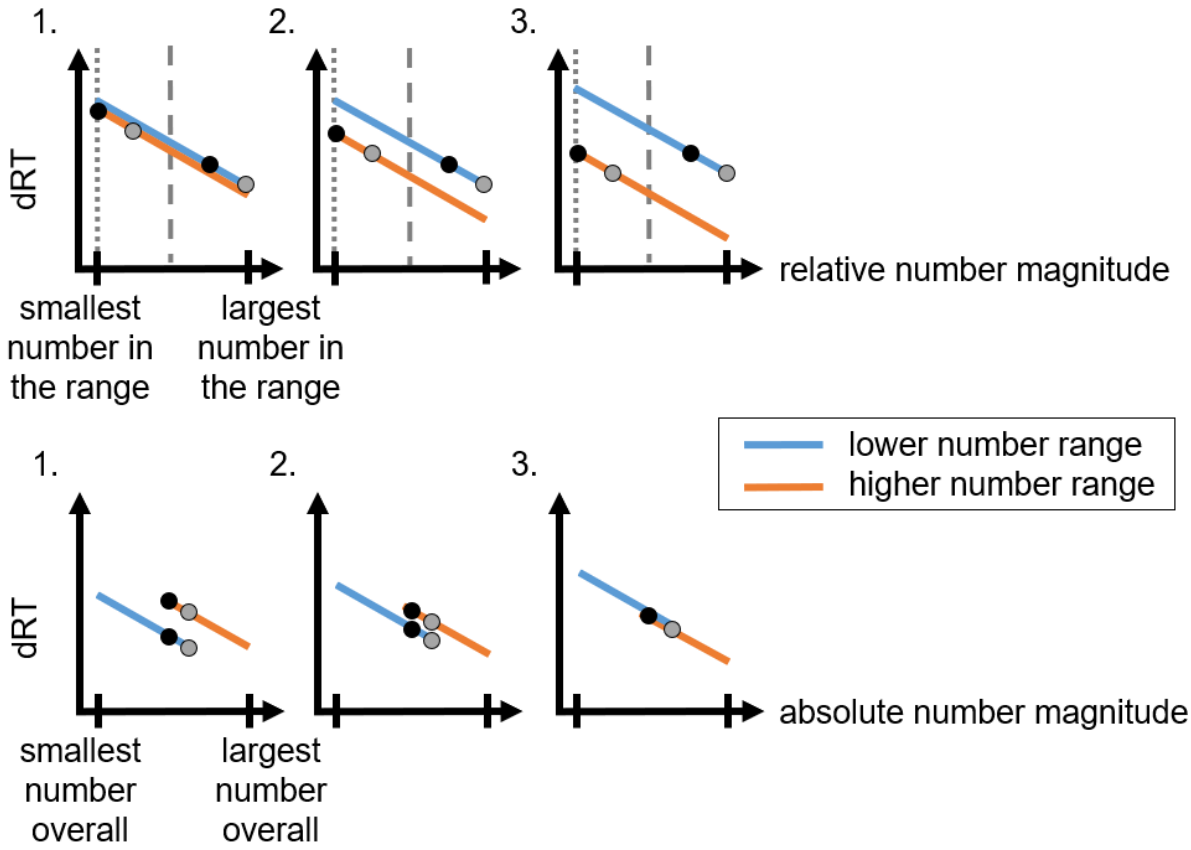
Possible scenarios of RMdependency and AMdependency

In the following, we want to present six possible scenarios regarding RMdependency and AMdependency of both the number mapping on the MNL and the strength of the SNARC effect. Apart from the regression slope that quantifies the strength of the SNARC effect, the smallest-number intercept (when relative magnitude of the numbers in both ranges is matched, i.e., the predicted dRT for 0 and 4 in Experiment 1 and for 1 and 4 in Experiment 2) and the mean-number intercept (i.e., the predicted dRT for 2.5 and 6.5 in Experiment 1 and for 3 and 6 in Experiment 2) can be determined in order to investigate the number mapping on the MNL. When discussing RMdependency and AMdependency of the SNARC effect, the following scenarios are possible (see Figures S1 and S2 and Table S1):

1. RMdependency of the number mapping on the MNL, but no difference in the strength of the SNARC effect between number ranges (i.e., different dRTs of critical numbers that are part of both number ranges, namely 4 and 5)
2. Both RMdependency and AMdependency of the number mapping on the MNL, but no difference in the strength of the SNARC effect between number ranges (i.e., different dRTs of critical numbers, different smallest-number intercepts, and different mean-number intercepts)
3. AMdependency of the number mapping on the MNL, but no difference in the strength of the SNARC effect between number ranges (i.e., different smallest-number intercepts and different mean-number intercepts) – note that concluding RMdependency of the number mapping on the MNL from finding a significant SNARC effect in both number ranges without testing dRTs of critical numbers is incorrect

Figure S1

Possible scenarios of RMdependency and AMdependency of the number mapping on the MNL

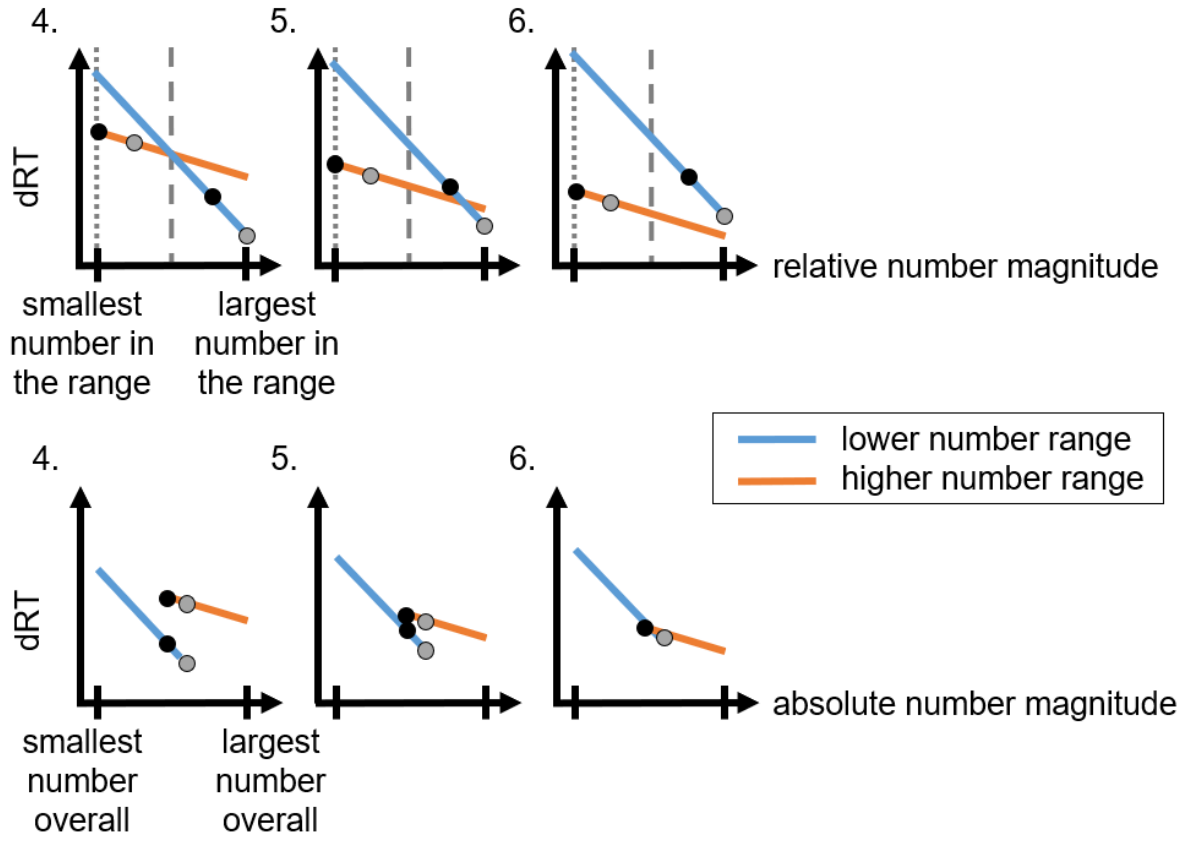


Note. This figure (retrieved from https://doi.org/10.17605/OSF.IO/Z43PM) illustrates Scenarios 1, 2, and 3, with the regression lines for the lower and higher number ranges being represented in blue and orange, respectively. In the upper part of the figure, relative number magnitudes are used for the x-axis, so that the regression lines for both number ranges start at their smallest and end at their largest number magnitude. For example, in Experiment 1, the dRTs for 0 (smallest number in the lower number range) and 4 (smallest number in the higher number range) are on the very left, and the dRTs for 5 (largest number in the lower number range) and 9 (largest number in the higher number range) are on the very right. In the lower part of the figure, the same scenarios are illustrated, but absolute number magnitudes are used for the x-axis. In our study, the absolute number magnitudes will be 0 to 5 and 4 to 9 in Experiment 1, and 1 to 5 (excluding 3) and 4 to 8 (excluding 6) in Experiment 2. For example, the dRTs for numbers 4 and 5 are on the very same spot of the x-axis for both the lower and the higher range, because they have the same absolute magnitude. The dotted line in the upper part of the figure depicts the intercept for the smallest number magnitude, and the dashed line depicts the intercept for the mean number magnitude in the respective number range. The black and the gray dots indicate the critical numbers being part of both the lower and the higher number range (i.e., 4 and 5).

4. AMdependency of the strength of the SNARC effect, and RMdependency of the number mapping on the MNL (i.e., different SNARC slopes, different dRTs of critical numbers, different smallest-number intercepts), as in Fias et al. (1996)
5. AMdependency of the strength of the SNARC effect, and both RMdependency and AMdependency of the number mapping on the MNL (i.e., different SNARC slopes, different dRTs of critical numbers, different smallest-number intercepts, and mean-number intercepts), as in Dehaene et al. (1993)
6. AMdependency of the strength of the SNARC effect and of the number mapping on the MNL (i.e., different SNARC slopes, different smallest-number intercepts, and different mean-number intercepts)

Figure S2

Possible scenarios of RMdependency and AMdependency of the strength of the SNARC Effect



Note. This figure (retrieved from https://doi.org/10.17605/OSF.IO/Z43PM) illustrates Scenarios 4, 5, and 6. For an explanation of magnitudes on the x-axis as well as concrete examples for data points, see *Note* of Figure S1.

Table S1

Possible Scenarios of RMdependency and AMdependency of the SNARC Effect

	Scenario					
Characteristic of the scenario	1	2	3	4	5	6
SNARC effect in both ranges	yes	yes	yes	yes	yes	yes
Different dRTs for critical numbers (4 and 5)	yes	yes	no	yes	yes	no
Different smallest-number intercept	no	yes	yes	yes	yes	yes
Different mean-number intercept	no	yes	yes	no	yes	yes
Different SNARC slopes	no	no	no	yes	yes	yes

Note. This table summarizes the characteristics of the six possible scenarios of RMdependency and AMdependency of the SNARC effect, which are described above and illustrated in Figures S1 and S2. The crucial distinction consists in whether dRTs, intercepts and slopes differ between the two ranges in both experiments.

The mean-number intercept that is illustrated by a dashed vertical line in Figures S1 and S2 helps distinguish the scenarios from each other. However, as can be seen in Table S1, it is not necessary to test it against zero in a Bayesian one-sample *t*-test, because the scenarios can be distinguished with the other tests. We expected to observe Scenarios 4 or 5 (for reasons, see main manuscript).

Response times in Experiment 1

The average RTs in ms per number in the two ranges from 0 to 5 and from 4 to 9 are plotted separately per response hand in Figure S3, and together for both response hands with standard errors in Figure S4.

Figure S3

Average RT in ms per number and per response hand (left: squares, right: triangles) in each range (low: blue, high: orange) in Experiment 1

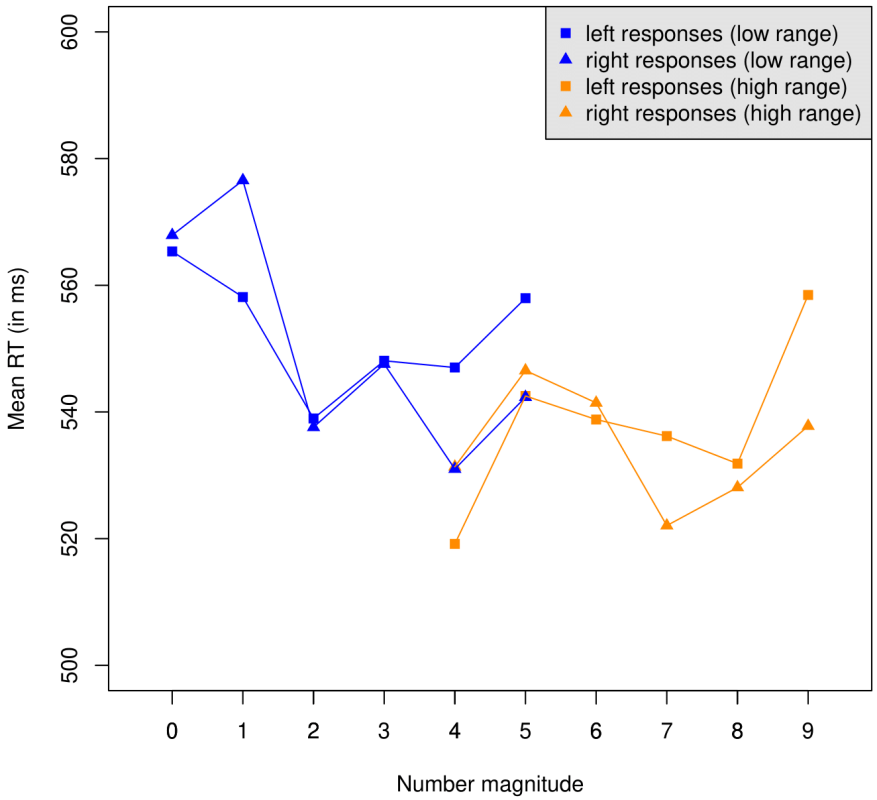
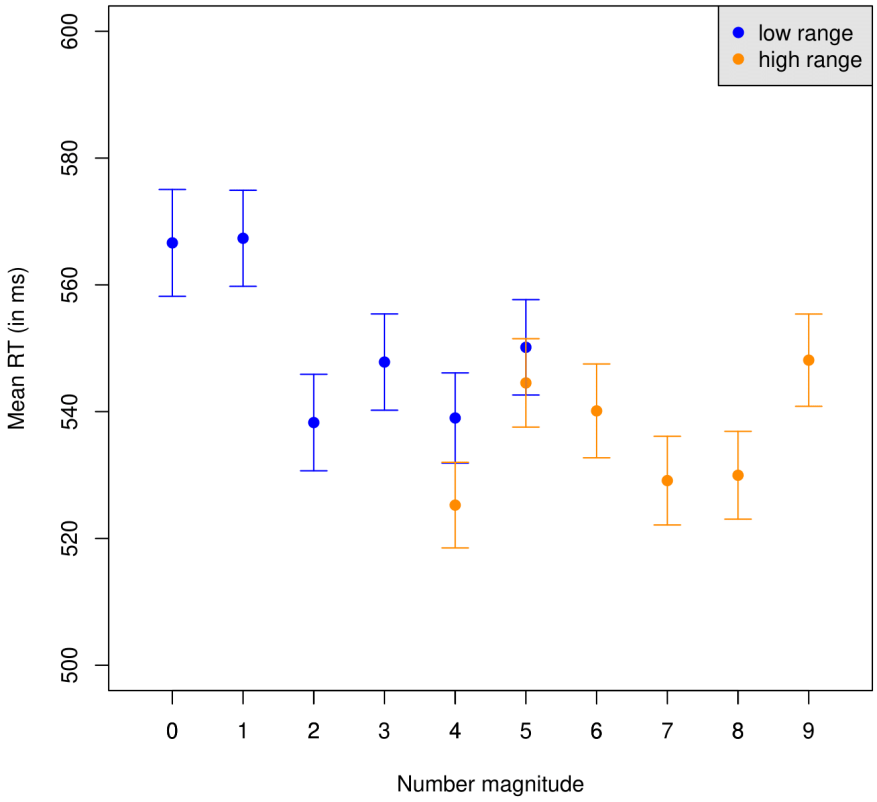


Figure S4

Average RT in ms (with with error bars indicating the respective SE) per number averaged over response hands in each range (low: blue, high: orange) in Experiment 1



Response times in Experiment 2

The average RTs in ms per number in the two ranges from 1 to 5 (excluding 3) and from 4 to 8 (excluding 6) are plotted separately per response hand in Figure S5, and together for both response hands with standard errors in Figure S6.

Figure S5

Average RT in ms per number and per response hand (left: squares, right: triangles) in each range (low: blue, high: orange) in Experiment 2

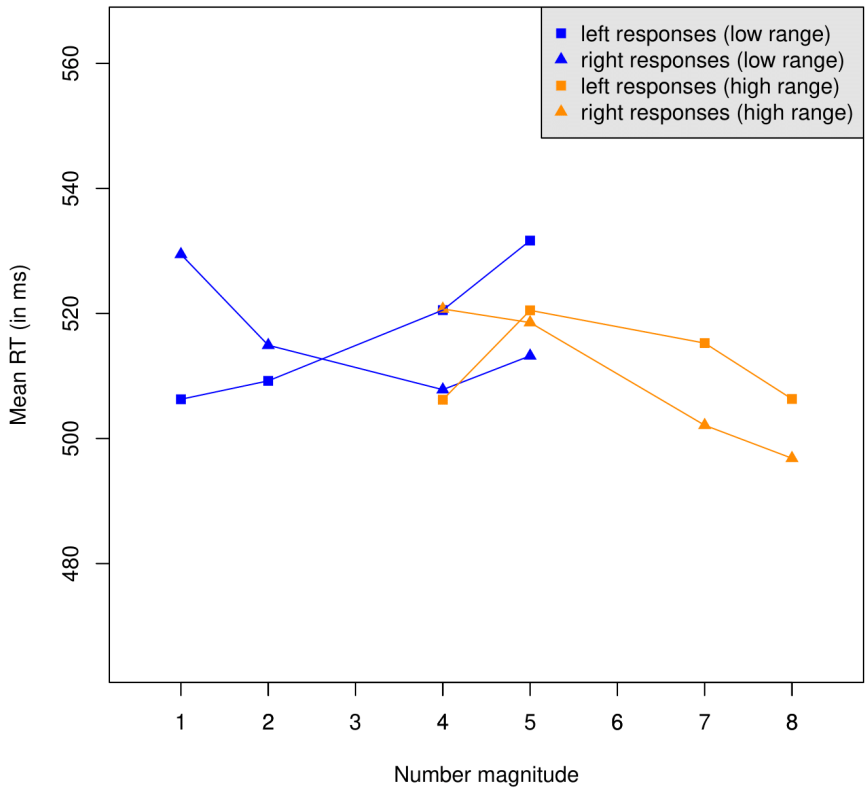
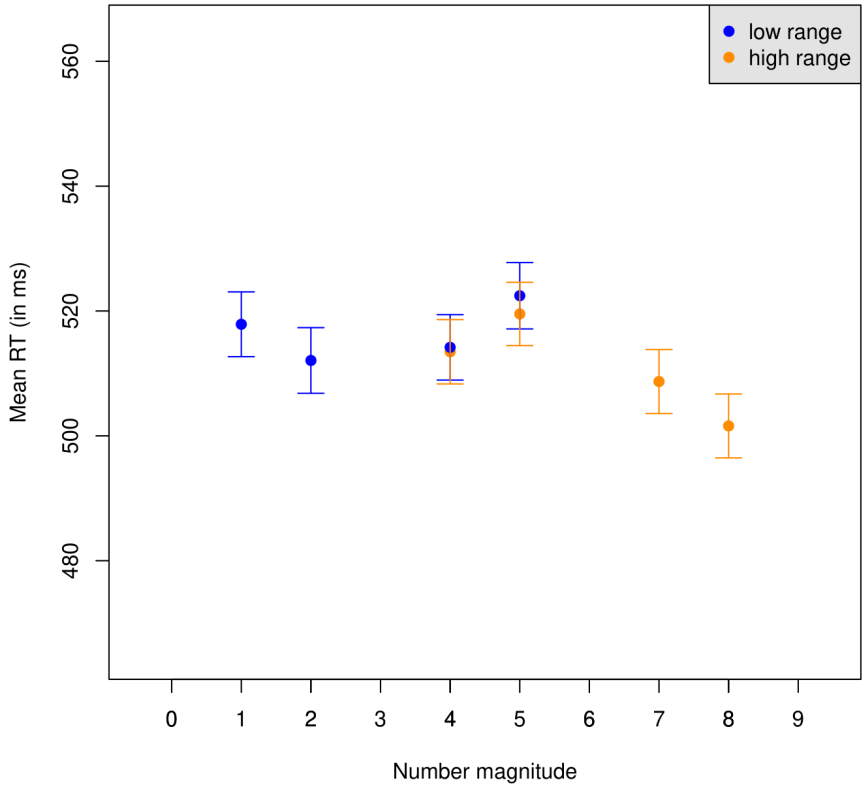


Figure S6

Average RT in ms (with error bars indicating the respective SE) per number averaged over response hands in each range (low: blue, high: orange) in Experiment 2



A6. Study 6

A6.1. Publication of Study 6

Citation:
Roth, L.*, Jordan, V.*, Schwarz, S., Willmes, K., Nuerk, H. C., van Dijck, J. P., & Cipora, K. (2024). Don’t SNARC me now! Intraindividual variability of cognitive phenomena – Insights from the Ironman paradigm. *Cognition, 248*, Article 105781. https://doi.org/10.1016/j.cognition.2024.105781

* shared first authorship

Copyright:
© 2024 Roth, Jordan, Schwarz, Willmes, Nuerk, van Dijck, Cipora (Open Access)
This article is published under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, provided the original authors and source are credited.

Cognition 248 (2024) 105781

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Don't SNARC me now! Intraindividual variability of cognitive phenomena – Insights from the Ironman paradigm

Lilly Roth [a,1], Verena Jordan [a,1], Stefania Schwarz [a], Klaus Willmes [b], Hans-Christoph Nuerk [a,c,d], Jean-Philippe van Dijck [e,f], Krzysztof Cipora [g,*]

[a] Department of Psychology, University of Tübingen, Germany
[b] Department of Neurology, University Clinic, RWTH Aachen University, Aachen, Germany
[c] LEAD Graduate School & Research Network, University of Tübingen, Germany
[d] German Center for Mental Health (DZPG), Germany
[e] Department of Applied Psychology, Thomas More University, Antwerp, Belgium
[f] Department of Experimental Psychology, Ghent University, Ghent, Belgium
[g] Centre for Mathematical Cognition, Loughborough University, United Kingdom

ARTICLE INFO

Keywords:
SNARC effect
Intraindividual stability
Reliability paradox
Ironman paradigm
Experimental psychology
Differential psychology

ABSTRACT

Two implicit generalizations are often made from group-level studies in cognitive experimental psychology and their common statistical analysis in the general linear model: (1) Group-level phenomena are assumed to be present in every participant with variations between participants being often treated as random error in data analyses; (2) phenomena are assumed to be stable over time. In this preregistered study, we investigated the validity of these generalizations in the commonly used parity judgment task. In the proposed Ironman paradigm, the intraindividual presence and stability of three popular numerical cognition effects were tested in 10 participants on 30 days: the SNARC (Spatial-Numerical Association of Response Codes, i.e., faster left−/right-sided responses to small/large magnitude numbers, respectively; Dehaene, Bossini, & Giraux, 1993), MARC (Linguistic Markedness of Response Codes; i.e., faster left−/right-sided responses to odd/even numbers, respectively; Nuerk, Iversen, & Willmes, 2004), and Odd (i.e., faster responses to even numbers; Hines, 1990) effects. We replicated the group-level effects; however, they were reliably present in only four to five (SNARC), six (MARC) or five (Odd) of 10 participants. Fluctuations seemed unsystematic, although the SNARC effect decreased over time along with reaction times. No correlation between the SNARC and MARC effects and sleep duration, tiredness, daytime, and consumption of stimulants were detected in most participants. These results challenge the frequent generalizations from group-level phenomena to individual participants and from single sessions to typical behavior. The innovative Ironman paradigm combined with bootstrap analyses permits unique insights into the intraindividual presence and stability of cognitive phenomena.

I behaved stubbornly, pursuing a semblance of order, when I should have known well that there is no order in the universe.

U. Eco, *The Name of the Rose.*

1. Introduction

Generalization is the core of inductive reasoning in science. By collecting several observations, supported by a toolkit of statistical inference, we generalize our findings beyond the observed data. Some of the generalizations we make are explicit, while some remain implicit and unspoken. In experimental psychology, the research process quite often goes from collecting data from a sample of individuals and searching for group-level effects to generalizing that the observed effects are present in the entire population by means of frequentist (or

* Corresponding author at: Centre for Mathematical Cognition, Loughborough University, United Kingdom.
E-mail addresses: lilly.roth@uni-tuebingen.de (L. Roth), Verena88@gmail.com (V. Jordan), stefania.marie.schwarz@gmail.com (S. Schwarz), willmes@neuropsych.rwth-aachen.de (K. Willmes), hc.nuerk@uni-tuebingen.de (H.-C. Nuerk), jean-philippe.vandijck@thomasmore.be (J.-P. van Dijck), k.cipora@lboro.ac.uk (K. Cipora).

[1] Equal contribution and shared first-authorship

https://doi.org/10.1016/j.cognition.2024.105781
Received 3 June 2023; Received in revised form 19 December 2023; Accepted 21 March 2024
0010-0277/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Bayesian) statistics. An extensive body of literature describes controversies and limitations of generalizability (Simons, Shoda, & Lindsay, 2017; Yarkoni, 2020). Extent and validity of sample-to-population generalizability is far from being resolved and remains beyond the scope of this paper.

Apart from this most obvious generalization, at least two more generalizations may be implicitly made by researchers at the same time. The first of these is the generalization that the group-level effect is reflected in each participant, which may be the case but is not necessarily warranted (see discussions on ecological fallacy in Fisher, Medaglia, & Jeronimus, 2018, and the group-to-person generalizability problem described by McManus, Young, & Sweetman, 2023). Until recently, this was even an explicit assumption in standard statistical methods, such as the general linear model. The general linear model assumes random variables that characterize the responses given by participants in the same experimental condition, which are independent and identically distributed. That is, these variables have the same mean and standard deviation, while any between-participant variability is treated as random error. The issue of between-participant variability is gaining more and more attention recently as new methods are being described to quantify individual prevalence of group-level psychological phenomena (e.g., Cipora et al., 2019; Haaf & Rouder, 2019; Rouder & Haaf, 2018; Rouder & Haaf, 2019; Zayas, Sridharan, Lee, & Shoda, 2019). Such studies usually demonstrate that not every participant, and in some cases even less than half of the participants, show a reliable and replicable group-level effect at an individual level. Even though the methods for quantifying individual prevalence are relatively new, the problem has a much longer history. Phenomena initially described as general principles of cognition, such as the advantage of global over local processing (Navon, 1977), turned out to reflect one of several possible cognitive styles (Happé & Frith, 2006; Poirel, Pineau, Jobard, & Mellet, 2008). In their dominance account, Rouder and Haaf (2018) differentiate between dominant phenomena like the congruency effect in the Stroop task (i.e., when naming ink colors of color words, responses are slower and more likely erroneous if the ink color and the word mismatch than if they match; Stroop, 1935), which are present in all people, and indominant phenomena like the advantage of global processing (Happé & Frith, 2006; Poirel et al., 2008), which are not present or even reversed in some individuals. Crucially, traditional significance testing at the group level cannot differentiate between dominant and indominant phenomena. For this, the presence of the phenomenon under scrutiny must be checked for each participant separately. A solution for estimating the proportion of the sample that reveals an observed group-level phenomenon at the individual level are bootstrap confidence intervals, which are built on the outcome measure's individual variability for each participant (Cipora, van Dijck, et al., 2019). Investigations of several cognitive phenomena have successfully employed bootstrap confidence intervals and shown them to be a suitable approach to answer the dominance question (e.g., Guida, Mosinski, Cipora, Mathy, & Noël, 2020; Hohol et al., 2020; Van Dijck, Abrahamse, & Fias, 2020; Van Dijck, Fias, & Cipora, 2022). In contrast to hierarchical modelling, which can estimate overall variability coming from separate sources (i.e., group level, participant level, and session level), bootstrap confidence intervals are a straightforward way to look at one level (i.e., participant level or session level) and disentangle the level's units (i.e., identify which participants or sessions show the cognitive phenomenon of interest reliably).

The second implicit generalization, which is our focus, is assuming that the effects observed at the participant level are stable over time and reflect their typical patterns of behavior. While interindividual consistency of group-level effects has become an issue recently, intraindividual consistency over time is largely overlooked in cognitive psychology. While in diagnostics or personality psychology the test-retest reliability of measured variables is a major topic, in cognitive psychology we usually measure only once and assume that basic cognitive effects are intraindividually stable. This generalization of a current effect to a permanent intraindividually stable effect seems to be mostly unchallenged in cognitive psychology, which might be due to the historical context of the development of scientific psychology.

1.1. Cognitive effects in correlational and experimental psychology

Historically, psychology has been divided between two approaches: correlational and experimental, as outlined by Cronbach (1957, 1975). Differential psychology focuses on explaining sources of variability between individuals, mostly taking the correlational approach. The goal of the measures employed is to capture maximum variability between participants, as observing variance is a necessary but not sufficient condition for observing covariance and in consequence correlation between measures. Experimental psychology, following the tradition of its founder, Wilhelm Wundt, is interested in general principles of human functioning. In contrast to differential psychology, experimental psychology uses measures and tasks that have been optimized to produce stable and replicable group-level effects. According to Wundt's reasoning, between-subjects variability reflects noise, which should be minimized (Jensen, 2006). However, designing measures and tasks with this goal and mindset makes the measures and tasks unsuitable for correlational studies (Hedge, Powell, & Sumner, 2018).

Measures used in contemporary cognitive psychology stem from experimental tradition. However, researchers use individual scores obtained by participants in a cognitive task as variables to be correlated with other variables or even as predictors for behavioral phenomena. Such correlations are quite often relatively low. In a seminal paper, Hedge et al. (2018) address this issue, referring to the reliability paradox. As explained above, cognitive tasks aim at producing stable group-level effects (in line with the experimental tradition) and are suited to minimize between-subjects variability. Thus, the largest part of the remaining between-subjects variability being observed when looking into individual scores can be assumed not to come from true differences between subjects regarding the measured construct but instead only from the measurement error. This may be the case, for instance, when the used task is too easy or when the sample is homogeneous regarding the measured effects (Paap & Sawi, 2016). Therefore, tasks which produce stable group-level effects often lack reliability as defined in differential psychology and psychometrics.

At the same time, there are numerous cognitive phenomena, which despite being robust at the group level, produce systematic between-subjects variability. One of them is the SNARC effect (Spatial-Numerical Association of Response Codes; Dehaene et al., 1993; Wood, Willmes, Nuerk, & Fischer, 2008, for a meta-analysis; Cipora, Soltanlou, Reips, & Nuerk, 2019, for an online replication; Toomarian & Hubbard, 2018, for a review), which describes a left/right side advantage in responding to small/large magnitude numbers, respectively. It is the hallmark effect in the field of numerical cognition and typically measured in a parity judgment task with a left and right response key, where the numerical magnitude affects the response pattern even though it remains task-irrelevant. The effect is indexed with slopes of regression, in which dRTs (mean RTs right side minus mean RTs left side) are regressed on the numerical magnitude for each participant separately. More negative values correspond to a more pronounced SNARC effect, that is, to a stronger association of small/large magnitude numbers with the left/right, respectively. Slopes are subsequently tested against zero with a one-sample *t*-test to check for the presence of the effect at the group level (repeated-measures regression approach by Lorch & Myers, 1990;

applied to the SNARC effect by Fias, Brysbaert, Geypens, & D'Ydewalle, 1996). Importantly, with this method, the SNARC effect is not tested at the participant level. Despite robustness at the group level, the SNARC effect varies between participants: Even descriptively, only between 70 and 80% of participants reveal negative slopes pointing towards a SNARC effect (Cipora, van Dijck, et al., 2019; Wood et al., 2008). When testing whether a reliable SNARC effect is present at the participant-level (i.e., unlikely to only reflect random error), the proportion of participants revealing it is even smaller and ranges from 35 to 45%. Relatively few individuals (5% to 10%) reveal the reversed effect (i.e., association of small/large magnitude numbers with the right/left side, respectively), and about half of the participants do not reveal the effect in any direction (Cipora, van Dijck, et al., 2019). The authors developed two bootstrapping approaches to determine confidence intervals for each participant to check whether their observed score is likely to occur by chance, which we make use of for the current study. Since the construction of these confidence intervals was non-parametric and based on each participant's individual behavioral pattern, it was possible to identify participants revealing a reliable regular or reversed effect. Specifically, the approach allows observed effects that might be a pure byproduct of random variability to be differentiated from (most probably) true underlying effects that would be unlikely to be produced only by fluctuations in reaction times. Some cognitive effects, such as the SNARC effect, reveal variability between individuals despite their robustness at the group level and would therefore be considered as indominant phenomena according to Rouder and Haaf (2018). However, presence of variability does not automatically imply reliability of the effect, because variability could be either produced by stable inter-individual differences (with high reliability as measured by test-retest correlations) or by high intraindividual variability of the effect (with low reliability as measured by test-retest correlations).

Using a design with enough trials increases the reliability of a task and of the estimated effects by decreasing the measurement error (Baker et al., 2021; Schuch, Philipp, Maulitz, & Koch, 2022; Zorowitz & Niv, 2023). In the case of the SNARC effect, enough trials (Cipora & Wood, 2017, for guidelines) ensure internal consistency as measured with the split-half reliability method (see Supplementary Materials in Cipora, van Dijck, et al., 2019). This seems to make the design promising for use in correlational studies (e.g., Cipora et al., 2016; Cipora & Nuerk, 2013). However, apart from robustness at the group level, sufficient between-subjects variability, and internal consistency, we should also be able to assume that the measure is stable over time, as it is only suitable to be correlated with other measures if it reflects a robust underlying mental process or structure. In terms of personality psychology, the question is whether a cognitive effect reflects a *trait* (i.e., long-term characteristic) or a *state* (i.e., transient, yet unstable cognitive process) of the participant.

From a theoretical point of view, the SNARC effect is meant to reflect the mind's tendency to map numerical magnitudes onto space. Many researchers hypothesize a cognitive construct in long-term memory called the *mental number line*, which might be helpful to process abstract concepts by using cognitive structures responsible for processing concrete objects (Dehaene, 2005). According to this hypothesis, the SNARC effect can be treated as a trait and correlated with other measures (similar claims of trait-like nature are also made about other cognitive phenomena, which researchers want to consider in correlational studies). For instance, a stronger mental number line (i.e., mapping numerical magnitude to space with greater ease and efficacy) might be associated with better mathematical performance. This translation from cognitive theory to individual differences is not very straightforward given that a correlation between the SNARC effect and mathematical performance is not consistently observed across studies (see Cipora, He, & Nuerk, 2020, for a review). A simple explanation for this is that the SNARC effect is not intraindividually stable over time.

Assuming intraindividual instability is not far-fetched given that the SNARC effect is susceptible to situated influences (Cipora, Patro, & Nuerk, 2018, for a review and taxonomy), such as variations of the experimental situation or task instructions. For instance, the SNARC effect's strength or even direction can be manipulated by previous tasks, such as reading texts in languages with left/right reading direction (Shaki & Fischer, 2008) or including small/large numbers near the left/right ends of the lines (Fischer, Mills, & Shaki, 2010). Moreover, representational manipulations can affect the SNARC effect, for instance by memorizing ascending or descending number sequences before parity judgment that need to be recalled later (Lindemann, Abolafia, Pratt, & Bekkering, 2008). Different response setups can also influence the strength or direction of the SNARC effect, for example, crossed hands in bimanual setups (Dehaene et al., 1993, but see Wood, Nuerk, & Willmes, 2006a), or verbal response setups with "left" and "right" labels instead of spatial response setups with left and right keys (Gevers et al., 2010). It is no contradiction to assume that the SNARC effect has both trait-like and state-like elements, because most personality models nowadays assume the expression of personality constructs (here: the mental number line) to be a function of both personality and situational context (e.g., Endler & Kocovski, 2001; Wilson, Thompson, & Vazire, 2017). The question which we aim to answer with the current study is more fundamental: If the situational context is not systematically manipulated and participants do the same SNARC experiment repeatedly, will their SNARC effect be stable? The intraindividual stability of the SNARC effect over time has not been established; it has never even been systematically investigated. This is a prerequisite for substantial aspects of our theory building about spatial-numerical associations and for correlational research conducted with the SNARC effect.

To sum up, the SNARC effect is a commonly investigated effect in numerical cognition that reveals large interindividual differences despite its high robustness at the group level. Moreover, it is highly susceptible to situated influences and its intraindividual stability over time remains unclear. Taken together, the SNARC effect seemed to be the perfect candidate for the present investigation of intraindividual variability.

1.2. Temporal stability of cognitive effects

Temporal stability can be quantified in terms of test-retest reliability and in terms of absolute stability of scores. Test-retest reliability measures the extent to which the order of participants' scores in a cognitive task remains consistent across measurement occasions (Zorowitz & Niv, 2023). More precisely, test-retest reliability is based on the correlation of scores obtained by the same participants in two different occasions and therefore provides information about the stability of individual scores in relation to the sample mean. However, it does not consider the stability of scores in absolute terms, that is their stability in units being used to measure it (e.g., in milliseconds). Test-retest reliability is perfect if all participants' values change by the same linear transformation (e.g., if all participants' mean RTs move by the same difference between the test and the retest), as the correlation coefficient is invariant to linear transformations. However, even if a similar group-level effect can be observed in the same sample on different occasions, it can come from non-correlated observations that largely differ in absolute terms. Importantly, the test-retest reliability is surprisingly low for some cognitive tasks and often not reported (Zorowitz & Niv, 2023). However, Parsons, Kruijt, and Fox (2019) argue that determining and reporting the reliability in cognitive psychology should become a standard practice, because it can avoid many years of research using measures

unsuited to individual differences research. Moreover, Schuch et al. (2022) investigated four common cognitive-control measures and found that these were stable and reliable when measured at the group level but do not show high split-half and test-retest reliability when assessing intraindividual differences as done in correlational approaches. In contrast to test-retest reliability, absolute stability considers how similar individuals' scores obtained from two different occasions are (e.g., how much mean reaction time [RT] varies between different experimental sessions). While test-retest reliability implies that inconsistencies are entirely due to the measurement error, it is important to consider absolute stability to account for variability caused by the participants. As outlined above, bootstrapping can provide valuable insights into absolute stability. Establishing both test-retest reliability and even more absolute stability is crucial to ensure that generalizations from single sessions to individuals' typical behavior are justified, but studies investigating this are surprisingly rare.

Test-retest reliability and absolute stability should therefore not be confused with one another. A cognitive effect or measure can be highly reliable while not being stable over time. For instance, if RTs of all participants increased by 3000 milliseconds between two time points, the correlation between them would be 1 and the effect or measure would be perfectly reliable despite low stability of the scores. At the same time, high absolute stability does not imply test-retest reliability. If all participants had very similar RTs and there was very small random variation between time points, the absolute stability would be very high, but the reliability would be very low (Miller & Ulrich, 2013). To conclude, test-retest reliability and absolute stability, while intuitively similar, provide complimentary but distinct types of information.

On top of test-retest reliability and absolute stability, measurement precision can also not be concluded from robust and replicable group-level effects and is independent from test-retest reliability and absolute stability. Luck (2019) defines precision as the extent to which one gets the same values when measuring the same thing at different occasions (or when using different subsets of trials from the same task, like when calculating split-half reliability). In speeded RT tasks, this would be the trial-to-trial variability in RT for a given condition within one participant. The lower this variability is, the more precise the measurement is. Measurement precision is determined within a session and does therefore not say anything about absolute stability and test-retest reliability between sessions.

One approach that takes both between-subjects and within-subjects variations into account are intraclass correlations (ICCs; Hedge et al., 2018; Shrout & Fleiss, 1979). Several forms of ICCs exist, but all of them are calculated as the ratio of the variance of interest over the total variance (i.e., the sum of between-subjects variance, between-sessions variance, and error variance). The different forms of ICCs can be distinguished depending on whether they are based on *consistency*, such that they consider the relative order of participants' scores across time or across raters, or based on *absolute agreement*, such that they consider the absolute stability of scores across time or across raters (Koo & Li, 2016). In other words, depending on the researchers' interest, ICCs can be calculated to put either the within-subjects variability of cognitive effects (e.g., their absolute stability between sessions) or their between-subjects variability (i.e., their robustness at the group-level) in proportion to the overall variance. However, ICCs are not routinely reported, and due to the different variants of this measure, they may be confusing to some readers.

Cognitive psychology does not usually investigate the intraindividual stability of effects over time, but only looks at average effects at the group-level, which are more robust with increasing sample sizes and more accurately captured with high measurement precision. Smith and Little (2018) argue that apart from addressing the replication crisis in psychology by increasing sample sizes, repeatedly measuring psychological phenomena in single individuals can also provide more reliable insights; citing Skinner: "It is more useful to study one animal for 1000 hours than to study 1000 animals for one hour" (Skinner, 1966, p. 21). Similarly, it is possible that situational factors affect the observed effects. Participants take part in studies at different times of the day, at different levels of tiredness, alertness, and motivation. While these fluctuations can be expected to cancel each other out at the group level, this is not the case at the participant level. Situational factors remain largely overlooked, even though they have been demonstrated to affect scores of cognitive tasks. For instance, noise during testing has been shown to affect measures of cognitive control (Hommel, Fischer, Colzato, van den Wildenberg, & Cellini, 2012).

There has been a lack of studies investigating the stability of cognitive phenomena over time, and only few studies have investigated test-retest reliability of the most established cognitive phenomena. Importantly, test-retest reliability has been shown to be low or moderate for the Simon, Stroop, Flanker, and other cognitive effects reflected by RT difference scores in the Simon, Stroop, Flanker, stop-signal, go/no-go, Posner cueing, Navon, anti-saccade, attentional network, and color-shape switching tasks; yet, test-retest reliability was high for RTs in these tasks (Cevada, Conde, Marques, & Deslandes, 2019; Eide, Kemp, Silberstein, Nathan, & Stough, 2002; Hedge et al., 2018; Paap et al., 2014; Paap & Sawi, 2016; Strauss, Allen, Jorgensen, & Cramer, 2005). While Paap et al. (2014) found a systematic decrease in the flanker effect over 100 sessions, the stability of other cognitive phenomena, to the best of our knowledge, has not been investigated over the course of many sessions and therefore especially their intraindividual stability remains unclear. Similarly, in the case of the SNARC effect, so far very few studies have addressed reliability, stability, and precision. The test-retest reliability of the SNARC effect in the parity judgment task has been found to be small to moderate ($r = 0.36$, $p = .01$ in Cipora & Göbel, 2013; $r = 0.41$, $p = .014$ in Georges, Hoffmann, & Schiltz, 2013; and $r = 0.274$, $p = .116$, or, after exclusion of three participants because of high Cook's distance, $r = 0.372$, $p = .04$, in Viarouge, Hubbard, & McCandliss, 2014). Note that effects with larger standardized effect sizes or with larger absolute sizes show higher reliability (Schuch et al., 2022). Further, Hedge et al. (2018) reported the between-subjects variance in the SNARC effect as proportion of overall variance to be rather low in a magnitude judgment task (ICC of 0.22). Thus, the intraindividual variability of the SNARC effect was relatively high compared to its interindividual variability. However, the 95% confidence interval (CI) for this ICC ranged from 0.00 to 0.49. Importantly, it remains unclear whether the relatively high intraindividual variability in the above-mentioned studies arose from systematic or unsystematic changes in the SNARC effect between sessions or if it was due to measurement imprecision. Crucially, to find out whether cognitive effects such as the SNARC effect are suitable for correlational (and not only experimental) psychological research, their within-subjects variability should be low and measurement precision should be high (Luck, 2019). Correlations between estimates of cognitive effects and other variables of interest as well as assessments of individual differences are limited by the test-retest reliability of the measures (Lord & Novick, 2008), and therefore only generalizable if both the cognitive effects and other variables are stable over time and if measurement errors are small (Schuch et al., 2022).

The mentioned limitations in reliability, absolute stability, and measurement precision in SNARC experiments have an important theoretical impact. In fact, we consider ease and high automaticity of mapping numbers to space to be an important mediator or metaphor for mathematical development (Cipora, Patro, & Nuerk, 2015), being at the core of much research about spatial-numerical associations (and the mental number line). If there is no such stable ability, but rather a

transient, highly fluent state of association of space and numbers, which bears not intraindividual stability, such conceptual implications would be void. Such considerations apply to other cognitive phenomena as well.

1.3. Objectives of the present study

This preregistered study aims at illustrating that the absolute stability of cognitive phenomena cannot be taken for granted. Specifically, we explored the intraindividual stability of the SNARC effect and other numerical cognition effects by checking whether observations in single experimental sessions reflect typical and stable behavior of the participants. For this purpose, we asked participants to perform the same parity judgment task on 30 out of 35 consecutive days. In the following, we will call the extensive study design where each participant completes the exact same experiment multiple times regularly within a given period the *Ironman paradigm* (in reference to a triathlon competition, as participation in the numerous sessions of our study likewise requires high perseverance). Furthermore, we investigated whether the effects vary systematically over time at the participant level or are associated with the participant's current state. In this study, we considered sleep duration, tiredness, consumption of stimulants, and time of the day as the most obvious situational factors.

As outlined above, the SNARC effect is highly replicable at the group level, even if reliably present in only less than half of the participants of previous studies (Cipora, van Dijck, et al., 2019). The SNARC effect is therefore an indominant phenomenon (as classified by Rouder & Haaf, 2018). Investigating its intraindividual stability was the focus of our study. Additionally, the SNARC effect is highly susceptible to situated influences (Cipora et al., 2018). The current study allows us to investigate its stability in individuals who reveal the typical effect, no effect, or the untypical reversed effect, and it potentially also permits us to investigate the role of situational context. Collecting data across 30 sessions per participant allowed us to clarify whether the SNARC effect indeed exists in every participant but was previously not reliably detectable in most samples due to lacking power, or whether it is genuinely present in less than half of the participants. Furthermore, although the SNARC effect was of our main interest, the parity judgment task allowed us to investigate other important numerical cognition effects as well. Namely, we also explored the reliable presence and intraindividual stability of the MARC effect (Linguistic Markedness of Response Codes; Nuerk et al., 2004; i.e., faster left−/right-sided responses to odd/even numbers, respectively) and of the Odd Effect (Hines, 1990; i.e., overall faster reactions to even than to odd numbers, irrespective of the response side). Finally, besides the mentioned numerical-cognition effects, we also investigated the intraindividual stability of RTs themselves. This measure should reveal less unsystematic variance within participants and instead more systematic changes over time as compared to difference scores (Cevada et al., 2019; Eide et al., 2002; Hedge et al., 2018; Paap et al., 2014; Paap & Sawi, 2016; Strauss et al., 2005), with participants becoming faster from session to session.

2. Method

The present study was preregistered (https://osf.io/p6u5h).

2.1. Participants

As preregistered, the sample size chosen for the study was small due to the high time requirement of the experiment. In total, 10 participants completed the task by conducting 30 sessions (one per day) of a bimanual parity judgment task within a 40-day period each (the preregistered period was 35 days, but three participants needed some additional days). Three authors of this study also participated in the experiment themselves. Additional participants were recruited through word of mouth at the University of Tübingen, Loughborough University, and Thomas More University of Applied Sciences, as well as in the personal environments of the authors. Students at the University of Tübingen were granted ten course credits. The study was conducted in accordance with the Declaration of Helsinki (World Medical Association, 2013). All subjects gave written informed consent. All participants completed the experiment in its entirety.

All participants were female, between 19 and 56 years old (M = 27.30 years, SD = 10.99 years), and left-to-right readers. Nine participants were right-handed and one was left-handed. Seven participants reported to be native German speakers (one of which bilingual with Spanish being the second native language), two were native Dutch speakers, and one was a native Italian speaker.

2.2. Materials

This study implemented a computerized speeded bimanual parity judgment task (task is available at https://osf.io/p6u5h). Participants were to decide the parity of a number presented on their computer screen by pressing a left key ("D") or a right key ("K") on their computer keyboard with their respective index finger. Stimuli (Arial font, 72 pt) were presented centrally in black (RGB: 0, 0, 0) against a gray background (RGB: 210, 210, 210). Each trial started with a 300 ms eye fixation cross (+) followed by a number (1, 2, 3, 4, 6, 7, 8 or 9) presented for a maximum of 2500 ms or until response. An inter-stimulus interval of 500 ms with a blank gray screen followed.

Each daily experimental session consisted of two parity judgment blocks with reversed response-to-key assignment. The block order was counterbalanced between days to minimize sequence effects. In each block, each number was presented 30 times leading to a daily total of 480 trials (8 numbers * 30 repetitions * 2 blocks). The presentation order was pseudo-randomized so that no number could appear more than twice in a row. Each block was preceded by a brief practice session (two trials per number), during which accuracy feedback was provided for incorrect responses by a red "X" (RGB: 255, 0, 0; Arial font, 72 pt., presented for 800 ms). The task was implemented in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012), version 3.3.10.

2.3. Procedure

Participants were asked to install the OpenSesame software on their computers. They received additional instructions on coding for participant and session numbers, the order of experiment versions to be used, and how to save and upload the data to a shared cloud storage. They were instructed to complete the task at any time during the day.

In each session, participants first completed a short questionnaire. They provided their participant code and current session number. Subsequently, they reported their sleep duration in the previous night (open input approximated to 0.5 h), their current level of tiredness (11-point Likert scale ranging from 0 [not tired at all] to 10 [almost falling asleep]), consumption of stimulants (coffee, cola, black tea) within one hour prior to the experiment (binary response format), as well as the current time of the day (three options: morning [waking up until 12 AM], afternoon [12 AM to 6 PM] or evening [6 PM until going to sleep]).

After completion of the questionnaire, instructions for the first practice session were provided, followed by the practice session.

Instructions for the first experimental block followed. Participants then completed the first experimental block. After the possibility to take a short break, the second block with reversed response-to-key assignment followed.

2.4. Data preparation

First, practice trials were discarded. The data was reviewed for any incomplete sessions (i.e., <480 completed trials) which were subsequently discarded from analysis. Furthermore, trials with incorrect responses were excluded from the data. Then, as preregistered, the algorithm proposed by Cipora, Soltanlou, et al. (2019) was applied for further data trimming. Trials with RTs under 200 ms were excluded as anticipations. Then, sequential trimming was used per participant and session. Trials with RTs above and below 3 SDs from the respective session's mean RT were excluded sequentially until means and SDs no longer changed. Finally, sessions with <66% remaining valid trials as well as sessions with empty number × response side cells configuration were fully excluded from analysis.

An additional inspection of the data showed that two participants (Participants 1 and 6) had not completed the experiment versions in the correct order (i.e., alternating execution of the two experiment versions to counterbalance block orders). Participant 1 used the same experiment version (right hand for odd numbers and left hand for even numbers in Block 1, right hand for even numbers and left hand for odd numbers in Block 2) for Sessions 5 through 25; Participant 6 used the same experiment version as above for Sessions 13 through 15). However, we included these two participants' data because this study focuses on intraindividual analyses and including their data did therefore not influence outcomes for other participants. Nevertheless, we explored whether deviating from the order of experiment versions specified in the instructions could have influenced the results.

2.5. Data analysis

Data were processed and analyzed with the statistical computing software R (R Core Team, 2022), using the R packages *cowplot* (Wilke, 2020), *dplyr* (Wickham, François, Henry, Müller, & Vaughan, 2023; Wickham, Vaughan, & Girlich, 2023), *GeneNet* (Schaefer, Opgen-Rhein, & Strimmer, 2021), *ggplot2* (Wickham, 2016), *lme4* (Bates, Maechler, Bolker, & Walker, 2015), *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2017), *plyr* (Wickham, 2011), *tidyr* (Wickham, François, et al., 2023, Wickham, Vaughan, & Girlich, 2023), *rmcorr* (Bakdash & Marusich, 2017). Raw data along with all analysis scripts are available at https://osf.io/p6u5h. Note that, because we had preregistered them for both effects, the following analyses will be applied to both the SNARC and MARC effects, and their results will be reported in an equally detailed way for both effects. However, the main focus of the current study was the SNARC effect.

2.5.1. SNARC and MARC slopes

To analyze the SNARC and MARC effects per participant and session, unstandardized and standardized SNARC and MARC slopes were calculated as proposed by Cipora, Soltanlou, et al. (2019). Since unstandardized slopes are more sensitive to variations in RT characteristics, fluctuations were expected to be more pronounced for unstandardized slopes than for standardized slopes. The standardized SNARC and MARC slopes should therefore better reflect the stable characteristics of an individual's spatial mapping of numbers (Lyons, Nuerk, & Ansari, 2015).

For calculating SNARC and MARC slopes, the individual regression method (Fias et al., 1996) was used as implemented in Cipora, Soltanlou, et al. (2019). dRTs were calculated for each participant, each session, and each number. These were regressed on number magnitude and contrast-coded parity (−0.5 and 0.5 for odd and even numbers, respectively). Both unstandardized and standardized regression slopes were calculated, with the latter being Fisher-z transformed to approximate normal distribution. Importantly, if the numbers 1, 2, 3, 4, 6, 7, 8, 9 are used, parity and magnitude predictors are orthogonal and presence of one predictor does not affect neither the unstandardized nor the standardized slope of the other. Both SNARC and MARC effects are reflected by negative slopes. Negative SNARC slopes reflect an increase of right- over left-hand advantage in milliseconds per increase in number magnitude of one unit, while negative MARC slopes reflect an increase of right- over left-hand advantage in milliseconds when the number parity status is even instead of odd. For each participant, the proportions of sessions revealing negative SNARC and MARC slopes were calculated as a first indication of the intraindividual stability of the effects. Furthermore, "grand" unstandardized and standardized SNARC and MARC slopes were calculated for each participant by aggregating all data across all sessions.

2.5.2. Intraindividual prevalence and stability of SNARC and MARC effects

Besides the calculation of the observed SNARC and MARC slopes, we were also interested in the reliability of the observed effects and the corresponding individual proportions of sessions revealing reliable SNARC and MARC effects. To test the reliabilities of all session and grand slopes, two bootstrapping methods were used to calculate individual 80%, 90%, 95%, and 99% CIs as proposed by Cipora, van Dijck, et al. (2019). Bootstrap CIs are calculated in a non-parametric approach, so that no distribution assumptions need to be made (Bland & Altman, 2015). Their major advantage is that their width is individually calculated for each participant and session so that measurement precision (as defined by Luck, 2019) can be estimated without being influenced by other participants or sessions. Hence, bootstrap CIs are applicable no matter whether the sample is homogeneous or heterogeneous regarding the cognitive phenomenon under scrutiny. Note that this is not the case for correlation-based reliability approaches (e.g., test-retest reliability, split-half reliability, internal consistency), which largely depend on the sample's variance. The two approaches described in the following are therefore well suited for estimating the prevalence of reliable SNARC and MARC effects at the participant level or even at the session level within each participant.

In the *H1 bootstrapping approach*, 30 trials were randomly sampled with replacement from all valid trials within each participant × session × number × response side configuration when calculating CIs for session slopes and 900 trials (i.e., 30 repetitions across 30 sessions) within each participant × number × response side configuration when calculating CIs for grand slopes. Based on these sets of selected trials, unstandardized and standardized SNARC and MARC slopes were calculated. This procedure was repeated 5000 times. Subsequently, the session and grand CIs were built based on the ranges containing the mid 80%, 90%, 95%, and 99% of the bootstrapped slopes. Since CIs calculated based on the H1 bootstrapping approach are used to test the extent of variation of an effect depending on which trials are considered for averaging, any observed (reversed) SNARC or MARC effect can be considered reliable when the corresponding CI does not include 0. Furthermore, the visualization of the observed slopes along with the H1 bootstrapped CIs allows evaluation of the similarities of SNARC and MARC effect estimates across sessions for each participant. We also looked at the extent to which the session CIs overlapped with the corresponding grand CIs.

In the *H0 bootstrapping approach*, two samples of 30 trials were randomly sampled with replacement from valid trials within each participant × session × number configuration for session CIs and 900

Table 1
Unstandardized (upper part) and standardized (lower part) SNARC effect. Within each part, one row corresponds to one participant. The left part shows descriptive results: proportions of sessions with negative and positive slopes; mean slope (averaged slopes from each session), standard deviation, and grand slope (calculated by aggregating all trials across all sessions). The two following columns correspond to the CI bounds for grand slopes resulting from both bootstrapping approaches. The following part summarizes the proportions of sessions with reliable negative (regular), no reliable, and reliable positive (reversed) effects according to both bootstrapping approaches.

	Descriptives: Proportions [%]		Descriptives: Slopes			90% H1 CIs: Grand		90% H0 CIs: Grand		90% H1 CIs: Proportions of sessions [%]			90% H0 CIs: Proportions of sessions [%]		
Participant	Slopes < 0	Slopes > 0	Mean slope	SD (slope)	Grand slope	lower	upper	lower	upper	Reliable effect <0	No reliable effect	Reliable effect >0	Reliable effect <0	No reliable effect	Reliable effect >0
Unstandardized SNARC effect															
1	83.33	16.67	−5.13	4.77	−4.83	−5.71	−3.95	−0.87	0.87	63.33	36.67	0.00	60.00	40.00	0.00
2	56.67	43.33	−0.50	1.76	−0.49	−0.96	−0.01	−0.50	0.49	20.00	76.67	3.33	20.00	76.67	3.33
3	93.33	6.67	−2.70	2.24	−2.72	−3.30	−2.13	−0.61	0.61	50.00	50.00	0.00	46.67	53.33	0.00
4	53.33	46.67	−0.01	1.46	0.06	−0.43	0.55	−0.50	0.51	10.00	86.67	3.33	6.67	90.00	3.33
5	63.33	36.67	−1.35	2.68	−1.16	−1.82	−0.47	−0.66	0.69	26.67	73.33	0.00	26.67	73.33	0.00
6	51.72	48.28	−0.39	2.89	−0.24	−1.01	0.56	−0.78	0.79	10.34	82.76	6.90	10.34	82.76	6.90
7	33.33	66.67	0.44	2.04	0.73	0.06	1.39	−0.67	0.68	3.33	86.67	10.00	3.33	86.67	10.00
8	53.57	46.43	−0.73	4.31	−0.49	−1.75	0.83	−1.28	1.27	10.71	85.71	3.57	10.71	85.71	3.57
9[a]	70.00	30.00	−1.17	2.17	−1.04	−1.60	−0.48	−0.56	0.57	20.00	80.00	0.00	16.67	83.33	0.00
10	66.67	33.33	−0.56	2.09	−0.53	−1.28	0.22	−0.76	0.75	10.00	86.67	3.33	10.00	86.67	3.33
Standardized SNARC effect															
1	83.33	16.67	−0.58	0.61	−1.43	−1.69	−1.06	–	–	63.33	36.67	0.00	–	–	–
2	56.67	43.33	−0.11	0.39	−0.33	−0.61	0.00	–	–	20.00	76.67	3.33	–	–	–
3	93.33	6.67	−0.30	0.24	−0.37	−0.45	−0.29	–	–	50.00	50.00	0.00	–	–	–
4	53.33	46.67	0.02	0.38	0.02	−0.15	0.19	–	–	10.00	86.67	3.33	–	–	–
5	63.33	36.67	−0.17	0.35	−0.63	−0.94	−0.23	–	–	26.67	73.33	0.00	–	–	–
6	51.72	48.28	−0.09	0.43	−0.07	−0.29	0.16	–	–	10.34	82.76	6.90	–	–	–
7	33.33	66.67	0.08	0.23	0.14	0.01	0.26	–	–	3.33	86.67	10.00	–	–	–
8	53.57	46.43	−0.04	0.39	−0.23	−0.66	0.31	–	–	10.71	85.71	3.57	–	–	–
9[a]	70.00	30.00	−0.19	0.44	−0.73	−1.04	−0.30	–	–	20.00	80.00	0.00	–	–	–
10	66.67	33.33	−0.12	0.35	−0.16	−0.38	0.07	–	–	10.00	86.67	3.33	–	–	–

[a] Participant 9 was left-handed.

trials within each participant × number configuration for grand CIs. The two samples were then treated as "left side" and "right side" responses respectively to calculate unstandardized SNARC and MARC slopes (as discussed by Cipora, van Dijck, et al., 2019, the H0 bootstrapping approach is not suitable for standardized slopes).[2] Again, this procedure was repeated 5000 times. The ranges containing the mid 80%, 90%, 95%, and 99% of the bootstrapped slopes were subsequently used to build the corresponding CIs of slopes likely to be observed if there was no association between magnitude / parity and response side. Any observed slopes outside these intervals can therefore be considered as reliable (reversed) SNARC or MARC effects.

In the analysis, we focus on 90% CIs, as in Cipora, van Dijck, et al. (2019). Results for other confidence levels are presented in Appendix A.

2.5.3. Time-series analysis of SNARC and MARC effects

To test for autocorrelations within the observed SNARC and MARC slopes, time-series analyses were conducted for each participant. Therefore, correlations between time-series with different time lags of one, two, three, four, and five days were calculated. A lag-1 correlation represents the correlation between a slope on day X and its equivalent on day X + 1, and correspondingly, lag-2, lag-3, lag-4, and lag-5 correlations represent the correlations between a slope on day X and its equivalents on day X + 2, X + 3, X + 4, and X + 5, respectively. It was of interest whether there would be a monotone decrease in the autocorrelations with an increase in time lags (lag-1 larger than lag-2 larger than lag-3, etc.). For each participant, the permutation-based variant of the C statistic (Tryon, 1982) was used to test deviations from randomness. Furthermore, each of the five autocorrelations per participant was tested against a distribution of correlations obtained by randomly reshuffling the observations (without replacement) and repeating this procedure 5000 times. Any correlation outside the mid 95% of this distribution would be considered a significant autocorrelation.

2.5.4. Influences of predictor variables on SNARC and MARC effects

To obtain some insights on possible predictors of intraindividual SNARC and MARC effect fluctuations, we tested whether the magnitude of SNARC and MARC slopes in a single session were influenced by the participants' sleep duration during the previous night, their current level of tiredness, consumption of stimulants within one hour prior to the experiment or time of the day. Therefore, repeated-measures correlations of sleep duration, tiredness, and consumption of stimulants with the SNARC and MARC slopes were calculated as proposed by Bakdash and Marusich (2017) using their R package *rmcorr*. Repeated-measures correlation models are linear mixed models (LMMs) with a fixed linear predictor, which account for the within-subjects design with random participant intercepts. To test whether time of the day influences the SNARC and MARC effects, LMMs with time of the day as fixed categorical predictor and with random participant intercepts were computed by using the R package *lmerTest* (Kuznetsova et al., 2017). Repeated-measures correlations could not be used here because we did not expect time of the day to have an ordinal or even linear effect on SNARC or MARC slopes.

Additional repeated-measures correlations and LMMs were calculated to validate the predictor variables. As a plausibility check, the correlation between the arousal measures sleep duration and tiredness was calculated, which we expected to be negative. Moreover, we investigated whether sleep duration, tiredness, consumption of stimulants, and time of the day influenced the mean RTs and the standard deviations (SDs) of RTs revealed by the participants on the

[2] This is a deviation from the preregistration as while these data have been analyzed, Cipora et al. discovered a flaw in the reasoning for the H0 bootstrapping in case of standardized slopes. Therefore, in this work we also do not report this problematic analysis.

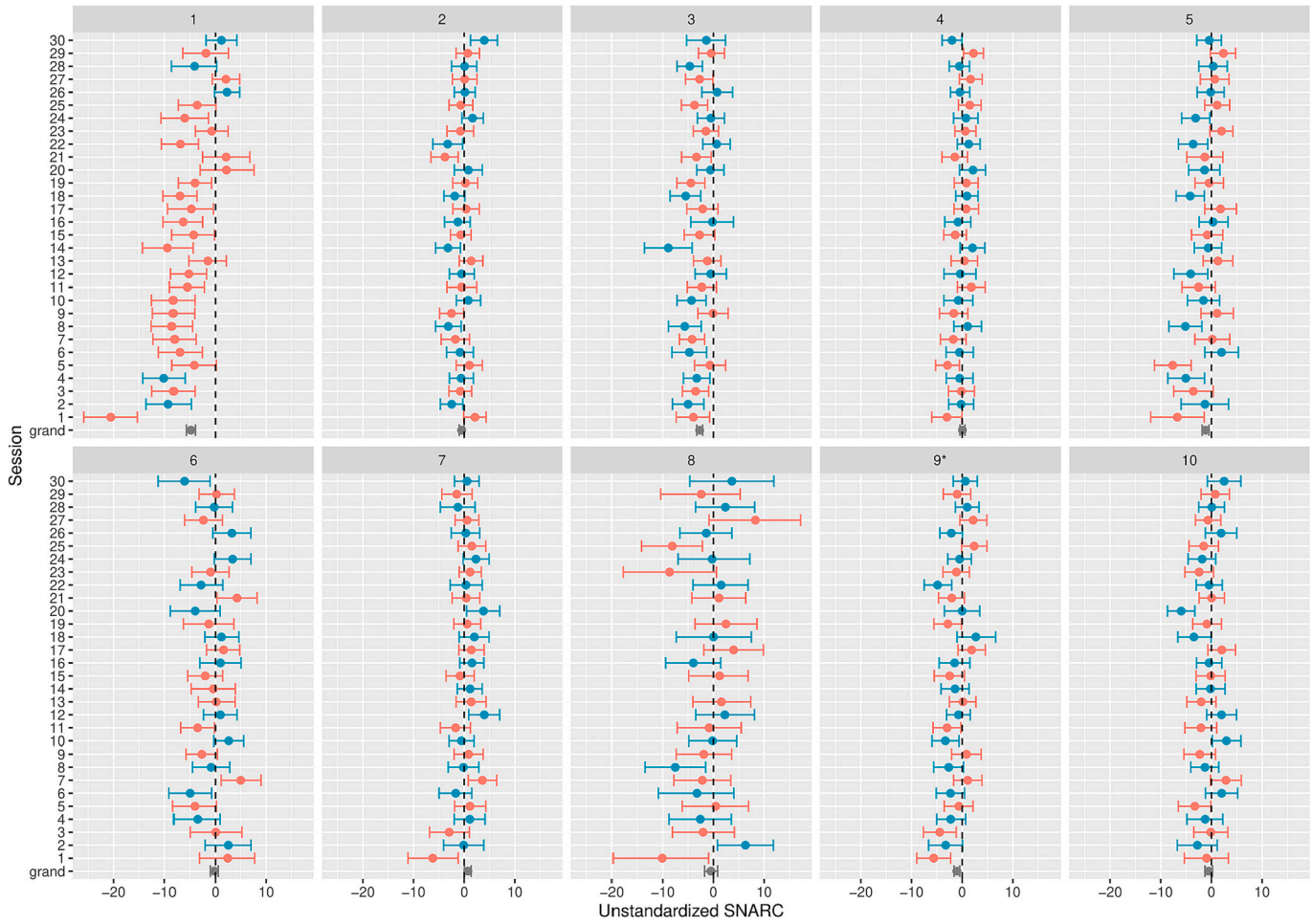


Fig. 1. Distributions of unstandardized observed grand and session SNARC slopes (dots) and 90% H1 CIs per participant.
Note. Colors represent the experiment version (red – Version 1: right hand for odd numbers and left hand for even numbers in Block 1, right hand for even numbers and left hand for odd numbers in Block 2; blue – Version 2: reversed order). *Participant 9 was left-handed.

corresponding days. A negative correlation between mean RTs and sleep duration, and between mean RTs and consumption of stimulants, as well as a positive correlation between mean RTs and tiredness would corroborate our measures. Additional analyses considering the Odd Effect, RT stability, session sampling and block order along with their justifications are reported in Appendices C to F. Finally, apart from investigating the systematicity of fluctuations in the SNARC and MARC effects via autocorrelations and Tryon's C statistic (Tryon, 1982), we also ran repeated-measures correlations between the SNARC and MARC slopes and the session number to investigate whether the two effects increase of decrease over time.

3. Results

3.1. Overview

After reviewing the data sets of all 300 sessions, data from one incomplete session (i.e., fewer than 480 trials) was excluded. The overall proportion of trials excluded due to incorrect responses was 7.0%. <0.001% of the trials were excluded because of anticipation. After applying the sequential trimming method as preregistered, another 3.3% of the trials were excluded. Two sessions were excluded due to excessive error rates (i.e., <66% valid trials). The assessments of a lack of valid RTs for any number × response side configuration did not lead to further session exclusions. Excluding the three sessions, 89.7% of all trials were retained for analysis.

3.2. Intraindividual prevalence and stability of SNARC effects

At an aggregated level (grand SNARC slopes), eight out of 10 participants descriptively revealed unstandardized and standardized SNARC effects (Participants 1, 2, 3, 5, 6, 8, 9, and 10). Among all 10 participants, the magnitudes of the grand slopes range from −4.83 to 0.73 for unstandardized SNARC effects and from −1.43 to 0.14 for standardized SNARC effects. When looking at the individual means of session slopes instead, the results remain similar, with the exception that one additional participant revealed a negative mean unstandardized SNARC slope (Participant 4). Without taking CIs into account, the intraindividual proportions of sessions showing SNARC effects range from 33.33% to 93.33%. Table 1 gives an overview of the results of the different analyses.

3.2.1. H1 bootstrapping approach

A more detailed look at the 90% CIs indicates that five out of 10

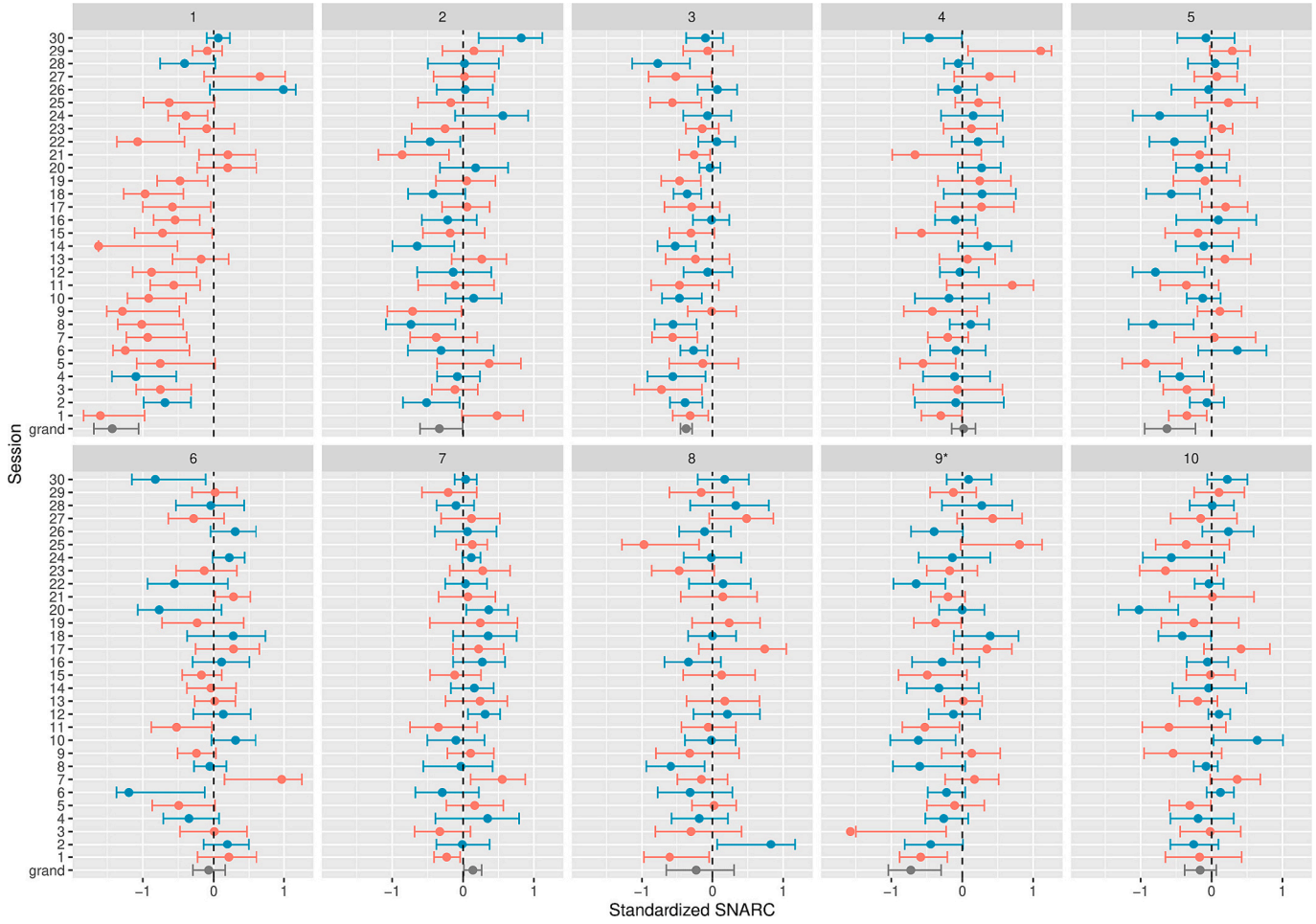


Fig. 2. Distributions of standardized observed grand and session SNARC slopes (dots) and 90% H1 CIs per participant.
Note. Colors represent the experiment version (red – Version 1: starting with right hand - odd numbers, left hand - even numbers (Block 1), and then reversed; blue – Version 2: reversed order). *Participant 9 was left-handed.

Table 2
Proportions of overlap between the individual session CIs and the corresponding grand CIs obtained from H1 bootstrapping for SNARC and MARC slopes.

Participant	Unstandardized SNARC effect	Standardized SNARC effect	Unstandardized MARC effect	Standardized MARC effect
1	76.67	46.67	63.33	73.33
2	93.33	96.67	80.00	83.33
3	93.33	86.67	46.67	50.00
4	96.67	100.00	43.33	40.00
5	86.67	80.00	50.00	70.00
6	93.10	100.00	58.62	44.83
7	96.67	96.67	43.33	30.00
8	96.43	100.00	67.86	89.29
9[a]	86.67	70.00	53.33	66.67
10	96.67	96.67	43.33	36.67

[a] Participant 9 was left-handed.

participants (Participants 1, 2, 3, 5, and 9) showed reliable unstandardized and standardized SNARC effects at the aggregated level (grand slopes). Participant 7 revealed reliable reversed effects for both measures. The intraindividual proportions of sessions revealing reliable effects ranged from 3.33% to 63.33% for both unstandardized and standardized SNARC effects. Again, the number of participants showing reliable SNARC effects in more than half of the sessions was low (one participant for unstandardized and standardized SNARC effects). Figs. 1 and 2 visualize the individual observed distributions of grand and session slopes with the corresponding 90% H1 CIs for unstandardized and standardized SNARC effects, respectively.

Note that the H1 CIs for standardized effects are not symmetrical around the observed effect estimates, in contrast to those for unstandardized effects. H1 CIs for negative estimates include less values that

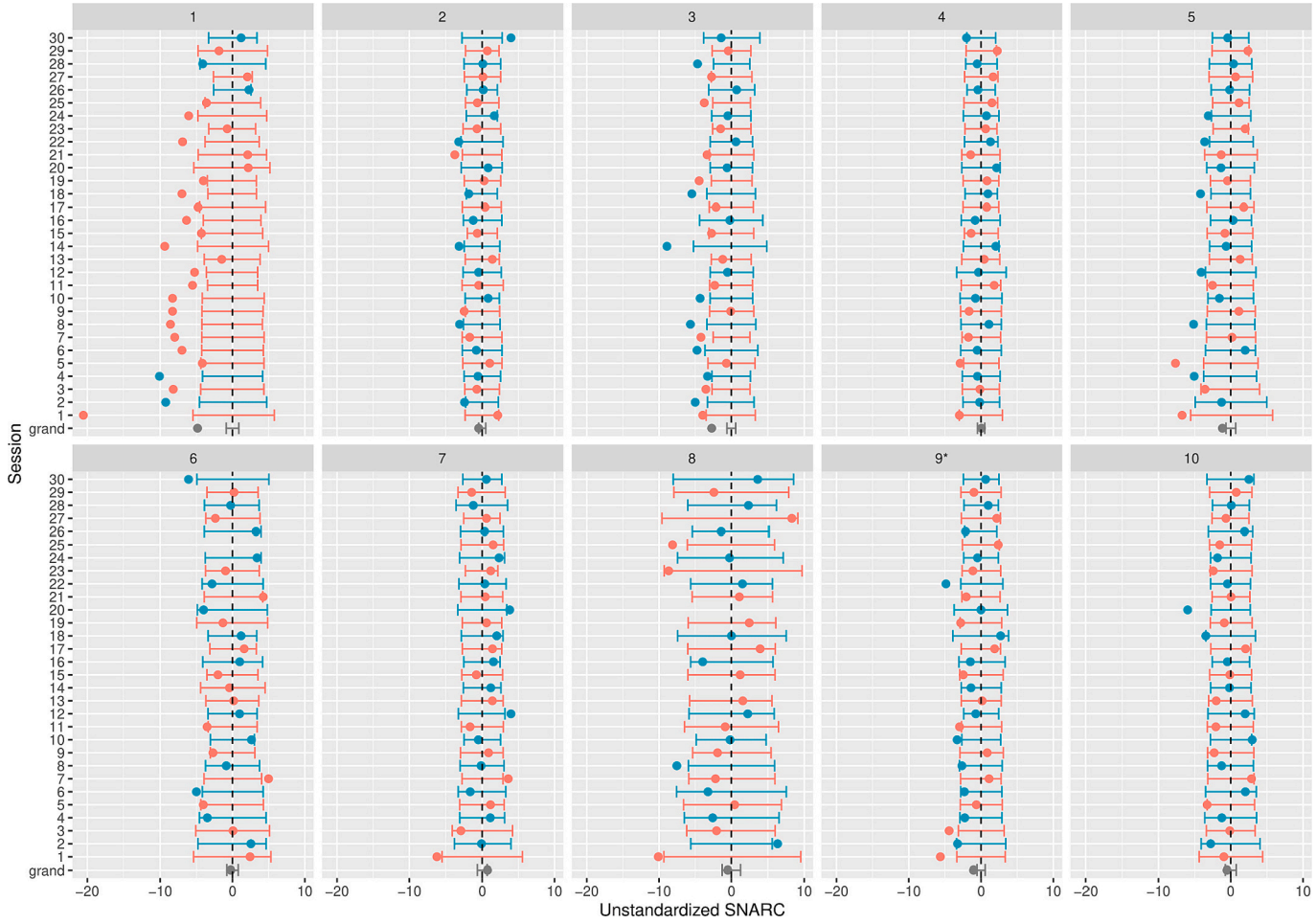


Fig. 3. Distributions of unstandardized observed grand and session SNARC slopes (dots) and 90% H0 CIs per participant.
Note. Colors represent the experiment version (red – Version 1: starting with right hand - odd numbers, left hand - even numbers (Block 1), and then reversed; blue – Version 2: reversed order). *Participant 9 was left-handed.

are smaller rather than larger compared to the estimate. Similarly, H1 CIs for positive estimates include less values that are larger rather than smaller compared to the estimate. The closer the estimates are to zero, the more symmetrical their H1 CI is. The reason for H1 CIs for standardized effects not being symmetrical is the restricted range of possible values. Namely, the minimum is −1 and the maximum is +1 (even though it is later Fisher-z transformed). Hence, the more extreme the estimate is, the less likely it is to obtain an even more extreme one by bootstrapping.

A closer look at Figs. 1 and 2 indicates that there is considerable intraindividual variation in the effect estimates across the different sessions and that the CIs of the different sessions do not overlap. Illustratively, the intraindividual proportions of session CIs that overlap with the grand CI were calculated. These proportions range from 76.67% to 96.67% for unstandardized SNARC effects and from 46.67% to 100% for standardized SNARC effects, as shown in Table 2. These intraindividual proportions are even lower for the overlap between session CIs and the grand slope estimate instead of the grand CI.

Looking at the H1 bootstrapping CIs for the grand slopes, one can notice that in case of unstandardized slopes, they are much narrower than those for single sessions (Fig. 1 for SNARC and Fig. 4 for MARC). This is because they consider many more observations, which leads to a more precise grand slope estimate compared to each session estimate. However, in case of standardized slopes, some CIs for the grand mean are as large as those for single sessions (Fig. 2 for SNARC and Fig. 5 for MARC). This is because the width of H1 bootstrap CIs for standardized slopes largely depends on the fit of the linear regression model rather than on the total number of trials. The model fit can vary more substantially depending on which points are being selected during bootstrapping.

For unstandardized slopes, H1 bootstrap CIs for the grand slope are roughly located at the mean of all session slopes (Fig. 1 for SNARC and Fig. 4 for MARC). Hence, most of them include zero or are relatively close to zero. However, in case of standardized slopes, some H1 bootstrap CIs for the grand mean are shifted towards extreme values (Fig. 2 for SNARC and Fig. 5 for MARC) and not necessarily located at the mean of all session slopes. We can observe such a pattern when the fit of the linear regression model to the grand dRTs is very high (see our plot with grand dRTs per participant per number in the data analysis folder on https://osf.io/p6u5h). For instance, the unstandardized grand SNARC slope is only slightly negative for Participant 9, but as the grand dRTs are well aligned to it, the CI for the standardized grand SNARC slope for Participant 9 is more negative compared to the CI for the standardized session SNARC slopes.

Table 3
Unstandardized (upper part) and standardized (lower part) MARC effect. Within each part, one row corresponds to one participant. The left part shows descriptive results: proportions of sessions with negative and positive slopes; mean slope (averaged slopes from each session), standard deviation, and grand slope (calculated by aggregating all trials across all sessions). The two following columns correspond to the CI bounds for grand slopes resulting from both bootstrapping approaches. The following part summarizes the proportions of sessions with reliable negative (regular), no reliable, and reliable positive (reversed) effects according to both bootstrapping approaches.

	Descriptives: Proportions [%]		Descriptives: Slopes			90% H1 CIs: Grand		90% H0 CIs: Grand		90% H1 CIs: Proportions of sessions [%]			90% H0 CIs: Proportions of sessions [%]		
Participant	Slopes < 0	Slopes > 0	Mean slope	SD (slope)	Grand slope	lower	upper	lower	upper	Reliable effect <0	No reliable effect	Reliable effect >0	Reliable effect <0	No reliable effect	Reliable effect >0
Unstandardized MARC effect															
1	73.33	26.67	−5.95	37.98	−5.30	−10.00	−0.46	−4.74	4.80	33.33	50.00	16.67	33.33	50.00	16.67
2	53.33	46.67	−0.67	15.00	−0.57	−3.21	2.19	−2.76	2.70	23.33	60.00	16.67	16.67	66.67	16.67
3	90.00	10.00	−39.48	30.90	−38.74	−42.01	−35.47	−3.28	3.34	80.00	13.33	6.67	80.00	13.33	6.67
4	66.67	33.33	−14.77	21.66	−14.76	−17.65	−11.92	−2.87	2.86	43.33	50.00	6.67	43.33	50.00	6.67
5	43.33	56.67	1.91	38.57	2.55	−1.31	6.28	−3.72	3.72	30.00	40.00	30.00	30.00	40.00	30.00
6	65.52	34.48	−15.64	34.79	−16.34	−20.87	−11.95	−4.45	4.45	37.93	48.28	13.79	37.93	48.28	13.79
7	76.67	23.33	−26.49	36.91	−27.14	−30.74	−23.61	−3.77	3.86	56.67	33.33	10.00	56.67	33.33	10.00
8	53.57	46.43	2.46	43.79	0.95	−6.09	7.83	−7.04	7.05	17.86	57.14	25.00	17.86	57.14	25.00
9[a]	40.00	60.00	0.90	26.43	0.64	−2.53	3.76	−3.09	3.10	26.67	50.00	23.33	26.67	50.00	23.33
10	70.00	30.00	−17.52	38.17	−17.55	−21.65	−13.45	−4.05	4.18	43.33	40.00	16.67	43.33	40.00	16.67
Standardized MARC effect															
1	73.33	26.67	−0.17	0.74	−0.18	−0.33	−0.02	–	–	33.33	50.00	16.67	–	–	–
2	53.33	46.67	−0.03	0.72	−0.07	−0.36	0.24	–	–	23.33	60.00	16.67	–	–	–
3	90.00	10.00	−0.93	0.72	−1.59	−1.77	−1.39	–	–	80.00	13.33	6.67	–	–	–
4	66.67	33.33	−0.58	0.86	−1.69	−1.96	−1.18	–	–	43.33	50.00	6.67	–	–	–
5	43.33	56.67	0.06	0.95	0.23	−0.11	0.52	–	–	30.00	40.00	30.00	–	–	–
6	65.52	34.48	−0.38	0.78	−1.42	−1.70	−0.89	–	–	37.93	48.28	13.79	–	–	–
7	76.67	23.33	−0.61	0.83	−1.70	−1.96	−1.33	–	–	56.67	33.33	10.00	–	–	–
8	53.57	46.43	0.04	0.71	0.08	−0.42	0.54	–	–	17.86	57.14	25.00	–	–	–
9[a]	40.00	60.00	0.10	1.02	0.07	−0.26	0.37	–	–	26.67	50.00	23.33	–	–	–
10	70.00	30.00	−0.48	1.06	−2.17	−2.24	−1.22	–	–	43.33	40.00	16.67	–	–	–

[a] Participant 9 was left-handed.

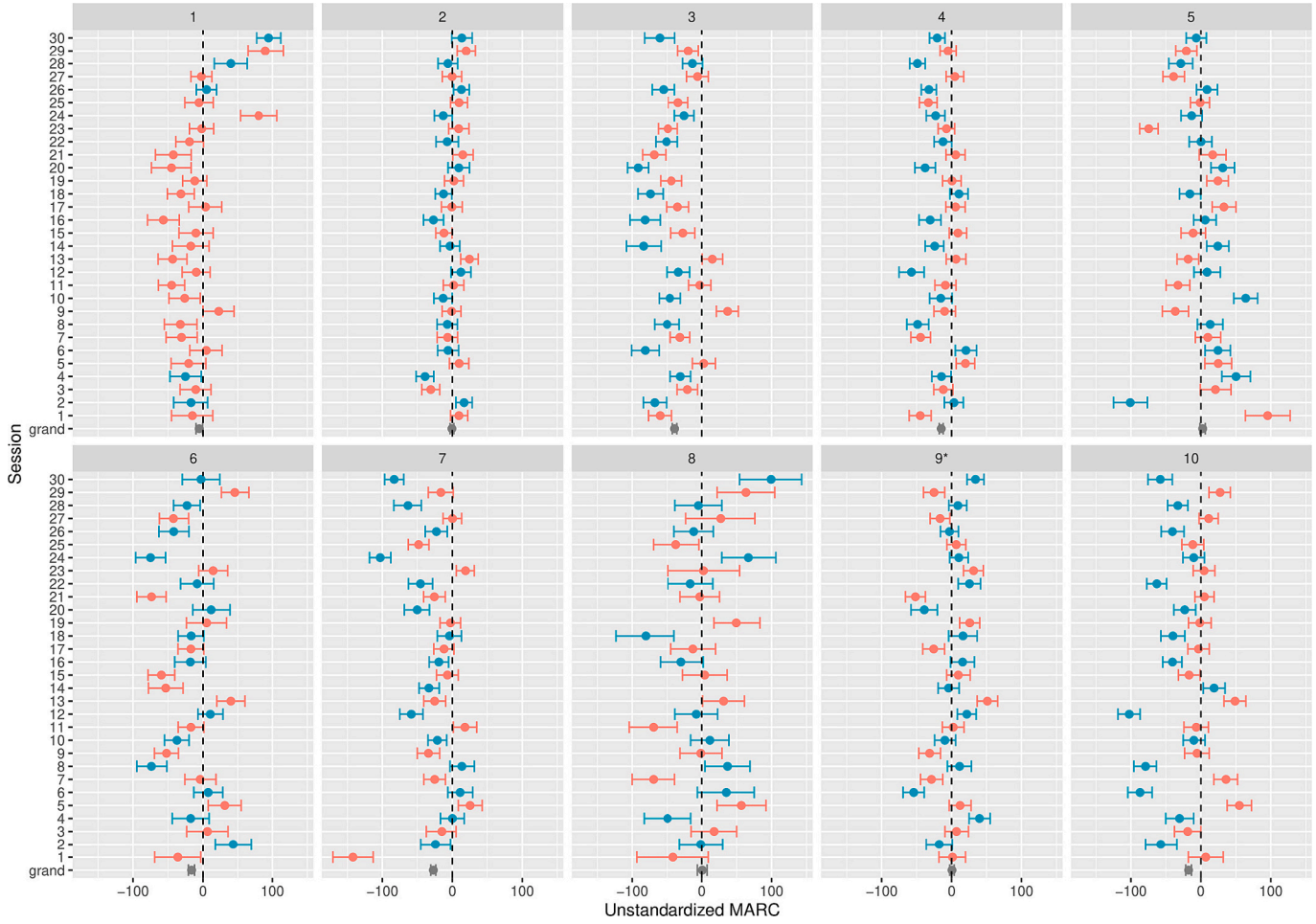


Fig. 4. Distributions of unstandardized observed grand and session MARC slopes (dots) and 90% H1 CIs per participant.
Note. Colors represent the experiment version (red – Version 1: starting with right hand - odd numbers, left hand - even numbers (Block 1), and then reversed; blue – Version 2: reversed order). *Participant 9 was left-handed.

3.2.2. *H0 bootstrapping approach*

At a 90% confidence level, four out of 10 participants (Participants 1, 3, 5, and 9) showed reliable unstandardized SNARC effects at the aggregated level (grand slopes), while Participant 7 showed a reliable reversed unstandardized grand SNARC effect. The intraindividual proportions of sessions revealing reliable effects ranged from 3.33% to 60.00%. Only Participant 1 revealed a reliable unstandardized SNARC effect in more than half of the sessions. As expected, the prevalence of reliable SNARC effects is therefore much lower when taking 90% CIs into account than the intraindividual proportions of descriptively negative slopes reported above. The intraindividual proportions of sessions revealing reliable reversed effects ranged from 0.00% to 10.00%. The individual distributions of observed grand and session unstandardized SNARC slopes including 90% H0 CIs are visualized in Fig. 3.

3.3. *Intraindividual prevalence and stability of MARC effects*

At an aggregated level (grand MARC slopes), seven out of 10 participants descriptively revealed negative MARC effect slopes. Among all 10 participants, the magnitudes of the grand slopes range from −38.74 to 2.55 for unstandardized MARC effects and from −1.70 to 0.23 for standardized MARC effects. The numbers of participants revealing effects remain the same when looking at the individual means of session slopes instead. Without taking CIs into account, the intraindividual proportions of sessions showing MARC effects range from 40.00% to 90.00% (see Table 3).

3.3.1. *H1 bootstrapping approach*

At the 90% confidence level, six participants (Participants 1, 3, 4, 6, 7, and 10) showed reliable unstandardized and standardized grand MARC effects, while no participant showed reliable reversed unstandardized and standardized grand MARC effects. For both unstandardized and standardized MARC effects, the intraindividual proportions of sessions revealing reliable effects ranged from 17.86% to 80.00% and two out of 10 participants (Participants 3 and 7) showed reliable MARC effects in more than half of the sessions. Figs. 4 and 5 visualize the individual observed distributions of grand and session slopes with the corresponding 90% H1 CIs for unstandardized and standardized MARC effects, respectively.

Again, the intraindividual proportions of session CIs that overlap with the grand CI were calculated to illustrate intraindividual variation in the effect estimates. These proportions range from 43.33% to 80.00% for unstandardized MARC effects and from 30.00% to 89.29% for standardized MARC effects (Table 2).

3.3.2. *H0 bootstrapping approach*

When looking at the 90% confidence level, six out of 10 participants (Participants 1, 3, 4, 6, 7, and 10) showed reliable unstandardized MARC effects at the aggregated level (grand slopes), while no participant showed reliable reversed unstandardized and standardized grand MARC effects. The intraindividual proportions of sessions revealing reliable effects ranged from 16.67% to 80.00%. Two participants (Participants 3 and 7) revealed a reliable unstandardized MARC effect in

Fig. 5. Distributions of standardized observed grand and session MARC slopes (dots) and 90% H1 CIs per participant.
Note. Colors represent the experiment version (red – Version 1: starting with right hand - odd numbers, left hand - even numbers (Block 1), and then reversed; blue – Version 2: reversed order). *Participant 9 was left-handed.

more than half of the sessions. The intraindividual proportions of sessions revealing reliable reversed effects ranged from 6.67% to 30.00%. The individual distributions of observed grand and session unstandardized MARC slopes including 90% H0 CIs are visualized in Fig. 6.

3.4. Time-series analysis of SNARC and MARC effects

The autocorrelations between the different observed SNARC and MARC slopes calculated for time series with time lags of one, two, three, four, and five days as well as the corresponding C statistic values are summarized in Table 4. Descriptively, none of the participants revealed a monotone decrease in the autocorrelations with an increase in time lags for any of the observed SNARC and MARC slopes. However, when looking at Tryon's C statistic, only Participant 1 revealed significant positive autocorrelations for unstandardized and standardized SNARC and MARC slopes, and only Participant 10 revealed significant negative autocorrelations for unstandardized and standardized MARC slopes.

A further investigation of each of the five time-lag autocorrelations by means of H0 bootstrapped 95% CIs reveals no significant autocorrelations for any of the participants between observed unstandardized SNARC slopes. For standardized SNARC slopes, only Participant 9 showed a significant negative lag-5 autocorrelation. Only Participant 10 revealed significant negative lag-1 autocorrelations for unstandardized and standardized MARC slopes. Additionally, Participant 2 showed a significant positive lag-2 autocorrelation for unstandardized MARC slopes and Participant 1 a significant positive lag-3 autocorrelation for standardized MARC slopes. All autocorrelations and bootstrapped CIs are visualized in Fig. B1 (unstandardized SNARC slopes), B2 (standardized SNARC slopes), B3 (unstandardized MARC slopes), and B4 (standardized MARC slopes) in Appendix B.

3.5. Influences of predictor variables on SNARC and MARC effects

Repeated-measures correlations were calculated to test whether fluctuations in SNARC and MARC slopes were predicted by the participants' sleep duration in the previous night, their current level of tiredness, and consumption of stimulants within one hour prior to the experiment. None of the calculated correlations was significant (see Table 5). Moreover, we computed LMMs with a categorical predictor to check whether time of the day influenced the SNARC and MARC effects, but these results were also non-significant (see Table 6). Note that only random participant intercepts were included within these repeated-measures correlations and LMMs for sleep, tiredness, stimulants, and daytime. LMMs for sleep, tiredness, and daytime that additionally include random participant slopes can be found in Appendix G. Due to the lack of sessions with prior consumption of stimulants (no session for Participants 5 and 9, only one session for Participant 7), no LMMs with random participant slopes were calculated for stimulants.

As preregistered, we conducted an additional analysis to cross-validate the measures. First, we calculated the repeated-measures correlation between sleep duration and current level of tiredness. This analysis revealed a modest correlation significantly different from zero

Fig. 6. Distributions of observed grand and session slopes (dots) and 90% H0 CIs per participant for unstandardized MARC effects.
Note. Colors represent the experiment version (red – Version 1: starting with right hand - odd numbers, left hand - even numbers (Block 1), and then reversed; blue – Version 2: reversed order). *Participant 9 was left-handed.

($r = -0.214$, $df = 285$, $p < .001$, 95% CI = [−0.322, −0.101]). As further validation measures, we calculated repeated-measures correlations and LMMs to investigate the relationship between all predictor variables (sleep duration, tiredness, consumption of stimulants, time of the day) and the mean RTs, as well as SDs of RTs that the participants revealed on the corresponding days. Most results remained non-significant (see Table 5 and Table 6). Only time of the day was a significant predictor of mean RTs, such that responses were approximately 13 ms faster when participants carried out the experiment in the morning than when they participated in the afternoon or evening (see Table 6).

3.6. Non-preregistered analyses

In addition to the preregistered analyses, we conducted non-preregistered analyses that can be found in the Appendix. Namely, we ran H0 bootstrapping for the intraindividual prevalence of the Odd Effect to be able to compare it with the individual prevalence of the SNARC and MARC effects (Appendix C); ran H1 bootstrapping for mean RTs, repeated-measures correlations of mean RTs with covariates, and time-series analyses for mean RTs (Appendix D); sampled group-level *t*-tests for the SNARC, MARC, and Odd effects (Appendix E); and used LMMs to test block order influences on the SNARC and MARC effects (Appendix F).

Further, we calculated repeated-measures correlations of the SNARC and MARC slopes, mean RTs, and SDs of RTs with the session number. Results indicated a significant decrease of the SNARC effect over time (unstandardized SNARC: $r = 0.319$, standardized SNARC: $r = 0.264$) but no significant change of the MARC effect over time was found (see Table 5). Session number correlated negatively with mean RTs, indicating responses getting faster over time, and with SDs of RTs, reflecting decreasing variability over time (see Table 5).

4. Discussion

In experimental psychology, researchers often draw conclusions from the robustness of effects at the group level to their robustness at the participant level (which is termed group-to-person generalizability problem by McManus et al., 2023); and they also assume that participant-level effects are stable over time and reflect the participant's typical behavior and thus a trait rather than a state characteristic. The goal of the current study was to provide a framework for evaluating these two types of generalizations. Combining the Ironman paradigm with bootstrapping techniques, we demonstrated that intraindividual stability of cognitive phenomena cannot be taken for granted.

As an exemplary phenomenon to illustrate this point, we chose the SNARC effect, which is one of the most thoroughly investigated phenomena in numerical cognition and highly replicable at the group level. It is typically investigated in a parity judgment task where other phenomena like the MARC effect and the Odd Effect can be observed as well. Hence, our aim was to analyze the intraindividual stability of the SNARC and other effects when looking at data from 10 participants with 30 sessions each within a 40-day period in the Ironman paradigm. Besides

Table 4
Autocorrelations in unstandardized and standardized SNARC and MARC effects with time lags of up to five days (lag-1, lag-2, lag-3, lag-4, lag-5) and Tryon's C statistic.

	Autocorrelations					Tryon's C statistic	
Participant	Lag-1	Lag-2	Lag-3	Lag-4	Lag-5	C	Z
Unstandardized SNARC effect							
1	0.080	0.190	0.042	−0.014	0.168	0.556	2.949*
2	0.047	0.229	−0.115	0.120	−0.286	0.215	1.140
3	−0.112	−0.104	−0.224	0.275	−0.028	−0.082	−0.435
4	−0.084	0.199	0.095	−0.020	0.012	−0.035	−0.184
5	−0.142	0.000	0.235	−0.203	−0.209	0.039	0.208
6	−0.090	−0.007	0.086	0.059	−0.108	−0.043	−0.222
7	0.052	0.150	−0.236	0.086	−0.083	0.170	0.902
8	0.062	0.057	−0.173	−0.129	−0.166	−0.018	−0.094
9[a]	0.044	−0.043	−0.173	0.218	−0.295	0.213	1.129
10	−0.060	−0.044	0.108	0.021	0.224	−0.114	−0.606
Standardized SNARC effect							
1	0.225	0.237	0.221	0.041	0.190	0.399	2.117*
2	0.055	0.292	−0.189	0.125	−0.293	0.168	0.893
3	−0.261	0.059	0.015	0.002	0.005	−0.035	−0.187
4	−0.118	0.052	−0.019	−0.214	0.046	−0.161	−0.853
5	−0.279	−0.093	0.228	−0.145	−0.234	−0.219	−1.162
6	−0.203	0.029	0.062	−0.088	−0.133	−0.137	−0.711
7	0.016	0.033	−0.031	0.072	−0.017	−0.039	−0.205
8	0.120	0.112	−0.071	−0.045	−0.132	−0.074	−0.377
9[a]	−0.047	−0.074	−0.204	0.209	−0.355	0.102	0.541
10	−0.065	−0.028	0.086	0.007	0.169	−0.127	−0.676
Unstandardized MARC effect							
1	0.135	0.312	0.188	−0.045	0.250	0.534	2.831*
2	−0.239	0.317	−0.313	−0.001	−0.155	0.187	0.993
3	−0.066	0.299	−0.140	0.183	0.044	−0.123	−0.650
4	−0.131	−0.020	−0.065	0.094	0.004	−0.155	−0.823
5	0.004	0.021	−0.180	0.194	0.038	−0.164	−0.869
6	0.210	−0.281	−0.140	−0.076	−0.086	0.139	0.723
7	−0.097	−0.142	0.101	0.121	−0.132	0.151	0.798
8	−0.278	0.128	0.113	−0.087	0.209	−0.013	−0.067
9[a]	0.221	−0.250	−0.218	−0.197	0.041	0.099	0.525
10	−0.529	0.102	−0.139	0.156	−0.077	−0.574	−3.042*
Standardized MARC effect							
1	0.133	0.276	0.293	−0.079	0.240	0.416	2.207*
2	−0.246	0.291	−0.317	0.062	−0.223	0.106	0.560
3	0.015	0.175	−0.072	0.240	−0.004	−0.060	−0.321
4	0.022	0.120	−0.203	0.109	0.128	−0.066	−0.348
5	−0.146	0.115	−0.207	0.221	0.035	−0.087	−0.463
6	0.238	−0.276	−0.168	−0.032	−0.037	0.196	1.022
7	−0.167	−0.150	−0.003	0.153	−0.217	0.000	−0.002
8	−0.246	−0.057	0.067	−0.119	0.269	−0.127	−0.647
9[a]	0.164	−0.134	−0.106	−0.275	0.030	0.108	0.572
10	−0.460	0.058	−0.166	0.178	−0.108	−0.518	−2.747*

[a] Participant 9 was left-handed.
* $p < .025$.

the calculation of the regular unstandardized and standardized SNARC and MARC slopes, as well as Odd Effect estimates at the session level and at the aggregated participant level, two bootstrapping approaches were used to obtain CIs for the grand slope to determine the reliability of the effects at the participant level. Moreover, we investigated the intra-individual stability of RTs themselves in a similar way. The data collected in the Ironman paradigm provides an interesting new perspective on the mentioned numerical-cognition phenomena.

The SNARC effect's robustness at the group level has been shown in various experimental tasks and for various populations (see Fischer & Shaki, 2014, for a review; Wood et al., 2008, for a meta-analysis). Interestingly, researchers have investigated its high interindividual variability found across studies (e.g., Wood, Nuerk, & Willmes, 2006b) but have not yet looked at its intraindividual stability over time. Instead, only the SNARC effect's test-retest reliability in terms of between-sessions correlations has been examined and found to be surprisingly low (Cipora & Göbel, 2013; Georges et al., 2013; Viarouge et al., 2014). Further, Hedge et al. (2018) have found that the SNARC effect's intra-individual variability is relatively high compared to its interindividual variability. Importantly, it remained unclear whether the low test-retest reliability and the relatively high intraindividual variability of the SNARC effect are due to systematic or unsystematic changes (i.e., lack of absolute stability) in the SNARC effect over time or possibly even due to measurement imprecision. This suggests that the SNARC effect is an indominant phenomenon (Rouder & Haaf, 2018), which is further supported by its susceptibility to situated influences (Cipora et al., 2018). In the current study, at an aggregated level over sessions, eight participants descriptively showed a grand SNARC effect, and seven participants descriptively showed a grand MARC effect. Even though these numbers must be interpreted with caution due to the very small sample size, it is in line with the proportions found in larger SNARC effect studies consistently reporting negative SNARC slopes for approximately 70% to 80% of participants (for an overview, see Cipora, van Dijck, et al., 2019). Interestingly, grand slopes were very small,

Table 5
Repeated-measures correlations between sleep duration, current level of tiredness, consumption of stimulants, and SNARC and MARC slopes, as well as RT characteristics.

Outcome	Predictors	*r*	*df*	*p*-value	95% CI
Unstandardized SNARC	Sleep	−0.028	285	0.637	−0.144, 0.088
	Tiredness	−0.046	286	0.436	−0.161, 0.070
	Stimulants	0.098	286	0.095	−0.018, 0.212
	Session	0.319	286	<0.001*	0.211, 0.419
Standardized SNARC	Sleep	−0.038	285	0.519	−0.154, 0.078
	Tiredness	−0.027	286	0.644	−0.143, 0.089
	Stimulants	0.016	286	0.787	−0.100, 0.132
	Session	0.264	286	<0.001*	0.154, 0.369
Unstandardized MARC	Sleep	−0.038	285	0.517	−0.154, 0.078
	Tiredness	−0.040	286	0.502	−0.155, 0.077
	Stimulants	0.042	286	0.481	−0.075, 0.157
	Session	0.043	286	0.466	−0.073, 0.158
Standardized MARC	Sleep	−0.084	285	0.155	−0.198, 0.032
	Tiredness	−0.077	286	0.192	−0.191, 0.039
	Stimulants	0.056	286	0.347	−0.061, 0.170
	Session	−0.001	286	0.984	−0.117, 0.114
Mean RT	Sleep	0.061	285	0.305	−0.055, 0.175
	Tiredness	0.001	286	0.987	−0.117, 0.115
	Stimulants	0.043	286	0.468	−0.073, 0.158
	Session	−0.540	286	<0.001*	−0.617, −0.452
SD RT	Sleep	−0.047	285	0.428	−0.162, 0.069
	Tiredness	0.043	286	0.463	−0.073, 0.158
	Stimulants	0.079	286	0.183	−0.037, 0.192
	Session	−0.248	286	<0.001*	−0.353, −0.136

* $p < .05$.

much smaller than SNARC slopes observed in single studies.

Expectedly, when taking CIs from the H0 and H1 bootstrapping approach into account, fewer participants show reliable effects, namely between four (H0 bootstrapping) and five (H1 bootstrapping) for the SNARC and six (H0 and H1 bootstrapping) for the MARC effect. A reversed SNARC effect was revealed by one and a reversed MARC effect by none of the participants, while the rest of the sample did not reveal effects in any direction. Participant 7, who revealed a reversed SNARC effect at the aggregated level, did not remarkably differ in any demographics from other participants: She was the third-youngest participant, female, right-handed, and bilingual (German and Spanish), but not proficient in any language with a right-to-left reading and writing direction. Again, the observed proportions for the SNARC measures are in line with the results by Cipora, van Dijck, et al. (2019) indicating a reliable SNARC effect in about 35% to 45% of participants. While at an aggregated level a reliable grand SNARC effect was shown by four to five participants, a reliable grand MARC effect by six participants, and a reliable grand Odd Effect by five out of 10 participants, the picture changes when looking at the intraindividual proportions of sessions with reliable effects. Only one participant showed reliable unstandardized and standardized SNARC effects in more than half of the sessions in both bootstrapping approaches. This means that even in most of the four to five participants who revealed reliable grand SNARC effects at the aggregated level, the effect could not be found on most days, indicating a very low intraindividual absolute stability of the SNARC effect. Crucially, we know that this is not an artifact originating solely from using a chronometric task, but instead a property specific to the SNARC effect. Hohol et al. (2020) have shown that, within the same datasets, different proportions of participants reveal the SNARC effect, the Numerical Distance effect, and the Numerical Size effect. Interestingly, the SNARC effect seems to be considered as a trait characteristic of individuals in the literature, but our results suggest that it is instead or concurrently a state characteristic. One might wonder how the only participant showing the SNARC effect in more than half of the sessions differs from others, who exhibited the SNARC effect only in few sessions or not at all. To start with, she ran the same experiment version (i.e., the same block order) for Sessions 5 through 25; however, we could not find any block-order influences on the SNARC effect (Appendix F). Moreover, this participant was the oldest, and the SNARC effect has been shown to increase with age (Ninaus et al., 2017). Also, she was the only participant who was entirely naïve to cognitive psychology experiments. However, to the best of our knowledge, there is no evidence for the SNARC effect to be affected by familiarity with cognitive psychology research, because it is an automatic effect and not deliberately produced by participants. Importantly, the participant had the second-largest mean RTs (cf. Fig. D1, Appendix D), and the SNARC effect is usually more pronounced in slower responses (e.g., Didino, Breil, & Knops, 2019). For the MARC and Odd effects, the number of participants showing reliable effects in most of the sessions was slightly higher. Two participants showed reliable unstandardized and standardized MARC effects and two participants showed a reliable Odd Effect in the majority of the sessions in the bootstrapping approaches. However, again, these numbers are considerably lower than expected, considering that six participants showed reliable grand unstandardized and standardized grand MARC effects and that five participants showed a reliable grand Odd Effect at the aggregated level.

The intraindividual instability of the effects over time is even more dramatic when taking into account that a considerable proportion of some participants' session CIs do not even overlap with their grand CI. Interestingly, the time-series analysis with autocorrelations and Tryon's C statistics suggest that the fluctuations in the SNARC and MARC effects between consecutive sessions are not systematic in most participants. Nevertheless, the SNARC effect but not the MARC effect decreased over sessions on average. Moreover, we did not find any relationship between four aspects of the participant's current state (sleep duration, tiredness, consumption of stimulants, and time of the day) and our measures of SNARC or MARC effects. Crucially, despite the strong variability and high instability over time in the SNARC and MARC effects and although the sample was so small, the group-level *t*-tests in our non-preregistered bootstrapping analysis (see Appendix E) with one randomly selected observed slope from each of our 10 participants revealed group-level SNARC and MARC effects in a quarter of all cases. Applying the same bootstrapping procedure for the Odd Effect shows that it would have been found in half of all cases. These findings emphasize that effects found at the group level in experimental cognitive psychology by testing random samples of participants only once do not necessarily reflect typical human behavior.One may ask why is this the case. At least three scenarios can potentially underlie the observed variability in the SNARC effect. First, the variability could be a purely statistical artifact, which is not related to underlying theoretical models. More precisely, Cipora, van Dijck, et al. (2019) have shown that the group-level effect can be driven by a minority revealing the effect in the expected direction (in the case of the SNARC effect, by around 40% of the individuals). At the same time, very few individuals reveal a reliable reverse effect. However, the fact that the SNARC effect gets induced from left to right rather than from right to left (even if still only in a minority of the sample) makes

Table 6
The influence of time of the day on unstandardized and standardized SNARC and MARC effects, on mean RTs, and on the SDs of RTs was investigated by comparing two LMMs in a Likelihood-Ratio test (LRT): LMM1 including only random intercepts per participant vs. LMM2 including both random intercepts per participant as well as a fixed term for time of the day, which was categorically coded with morning (M), afternoon (A), and evening (E). We report both the Akaike information criterion (AIC) and the Bayesian Information Criterion (BIC) for each model as well as the χ^2 statistic (two degrees of freedom) and the p-value resulting from the LRT. Moreover, our table displays estimates for the model of choice, namely variances of participant intercepts and of residuals as random effects, and slope estimates (with standard errors in brackets) for all three times of the day (LMM1) or separately in case of a significant LRT (LMM2) as fixed effects (dummy-coded with M as reference).

	LRT				Random effects		Fixed effects
	AIC	BIC	χ^2 (2)	*p*-value	Variance: participants	Variance: residuals	Slope estimates (*SE*)
Unstandardized SNARC effect							
LMM1	1486.8	1497.9			2.099	7.962	−1.211 (0.487)
LMM2	1488.0	1506.5	2.812	0.245			
Standardized SNARC effect							
LMM1	313.0	324.1			0.026	0.155	−0.150 (0.056)
LMM2	314.9	333.3	2.164	0.339			
Unstandardized MARC effect							
LMM1	2948.6	2959.7			137.350	1116.604	−11.542 (4.183)
LMM2	2950.9	2969.4	1.688	0.430			
Standardized MARC effect							
LMM1	767.0	778.0			0.086	0.721	−0.298 (0.105)
LMM2	769.5	788.0	1.451	0.484			
Mean RT							
LMM1	2918.8	2929.9					
LMM2	2919.1	2937.6	3.701	0.157	2421.350	906.392	M: 450.268 (15.849) A: 8.206 (4.383) E: 2.327 (4.419)
SD RT							
LMM1	2357.8	2368.9			422.039	138.165	66.303 (6.532)
LMM2	2361.7	2380.2	0.104	0.949			

this purely statistical explanation problematic.

Second, the SNARC effect might be present in everyone, but not precisely captured by the parity judgment task, so that both interindividual and intraindividual differences are just a by-product of the task. However, the explanation that the parity judgment task is imprecise or unsuitable for measuring the SNARC effect seems implausible, because its reliability is reasonable, especially in our Ironman paradigm with 900 data points per experimental cell for each participant. Moreover, this explanation seems unlikely in the light of results from Hohol et al. (2020) showing that variation in the SNARC effect is larger than in the Numerical Distance Effect and smaller than in the Numerical Size Effect (even though both other effects have no spatial component when it comes to lateralized responses). Their results show that cognitive effects differ in terms of their individual prevalence, even when calculated from the same data, the same participants, and the same task.

Third, there might be truly underlying trait differences in spatial-numerical associations. Some humans might associate numbers with space in a stable way, while others do not, which is in our view the most likely underlying scenario. This interindividual variability can occur on two levels: Individuals can differ in the spatial mental representations of numbers (consider the extreme differences between number-form synesthetes for an illustration) and in their intraindividual variability (i.e., susceptibility to situational context). This would mean that each individual has a personal underlying SNARC distribution with different means and standard deviations reflecting the likelihood of different effect sizes and directions. Therefore, some individuals show a regular and typically sized SNARC effect around −7 ms in most of the sessions (e.g., $SD = 1$ ms). In contrast, other individuals' SNARC effect might depend to a much larger degree on the situational context (e.g., on tasks preceding the parity judgment, on exact instructions, on aspects of the current state), causing stronger variations between sessions (e.g., mean slope of −7 ms, $SD = 5$ ms) and therefore sometimes inconsistent correlations. Several situational influences might influence the SNARC effect at the same time and even interact with each other. This would be in line with recent findings by Judd, Aristodemou, and Kievit (2023), who ruled out monocausal explanations for the effect sizes in eleven cognitive-control tasks when investigating variability in children tested within 25 sessions. They claim that variables such as attention, fatigue, and dopamine levels considered individually can at most only provide incomplete explanations for the high variability of cognitive effects or performances. Judd et al.'s results (Judd et al., 2023) showing that children with strong variability in some tasks only showed weak variability in other tasks speak in favor of this third potentially underlying scenario. A single dimension for variability across individuals and across tasks is therefore inadequate. What is more, a recent study by Belli and Fischer (2023) showing that breathing can shift spatial attention demonstrates that variability is present on a very short time scale. Specifically, processing of stimuli presented on the left/right was facilitated during exhalation/inhalation, respectively.

In contrast to the SNARC and MARC effects, which are based on differences in RTs between response sides, our additional non-preregistered analyses showed that RTs vary in a more systematic way (see Appendix D). We found positive autocorrelations and a trend over time towards faster responses in most participants. This is consistent with previous findings from other cognitive tasks (e.g., Cevada et al., 2019; Eide et al., 2002; Paap & Sawi, 2016; Strauss et al., 2005), where high test-retest reliabilities were found for RTs, although only low or moderate test-retest reliabilities were found for the actual effects of interest, such as the Simon, flanker, or Stroop effects. Along with responses getting on average faster over time, their variance also decreased on average over time (as reflected by negative correlations of the session number with mean RTs and SDs of RTs), reflecting that absolute stability (as defined in the Introduction) is low. Moreover, we replicated the tendency of smaller SNARC effects in faster responses (e.g., Gevers,

Verguts, Reynvoet, Caessens, & Fias, 2006). Also, in our attempt to replicate earlier observations regarding the influences of block order on the MARC effect in parity judgment (see Appendix F), we unexpectedly discovered the exact opposite pattern. While we found no influence of block order on the SNARC effect, interestingly, the MARC effect was stronger when the first block was MARC-compatible and the second block MARC-incompatible in the present study, while a previous analysis of datasets including over 1000 participants revealed a stronger effect in the reversed block order (Cipora et al., 2020). An explanation for the result of the previous study could be that participants who start with the MARC-incompatible block must both familiarize themselves with the task and overcome their natural response tendency in the first block; in the second block, they are already familiar with the task and do not have to counteract their spatial association of parity. Therefore, these participants have two reasons to be relatively slow in their first block and two reasons to be relatively fast in their second block, resulting in a rather strong MARC effect. Participants who start with the MARC-compatible block, however, must only familiarize themselves with the task but do not have to overcome their natural response tendency in the first block; in the second block, they must only counteract their spatial association of parity as they are already familiar with the task, which leads to a rather weak MARC effect. Importantly, while in that study block order differed between participants, it differed within participants in the present study, where we counterbalanced it between sessions. Our participants did not need to familiarize themselves with the task in each session's first block anew, such that the only difference between blocks was their MARC-compatibility. Hence, we can see that the influence of block order on the MARC effect seems to be different if familiarity with the task is ruled out.

Taken together, it seems that the 10 participants in the current study revealed response patterns that are rather typical in behavioral studies. By applying the Ironman paradigm with 30 sessions per participant, we collected enough data to grasp these basic response patterns. Hence, although future studies need to confirm this with larger samples, the high variability in SNARC and MARC effects that we have observed might not be due to our small specific sample; it might rather be a characteristic of the SNARC effect, MARC effect, and other cognitive phenomena.

4.1. Implications for different traditions in psychology

Taken together, these results may be a first indication that the SNARC and MARC effects are both characterized by the seemingly paradoxical attributes of high robustness in the population and concurrent low intraindividual robustness. Implications that we discuss for these effects can be applied to other cognitive phenomena, if these turn out to reveal high variability as well.

Importantly, the finding that the SNARC effect does not exist in every individual and is not stable within individuals does not seem to be grounded in measurement imprecision: The SNARC effect's reliability in the parity judgment task is typically within a reasonable range and not correlated with the proportion of participants showing a reliable SNARC effect, which was shown in an extensive reanalysis of 18 datasets with over 1000 participants by Cipora, Soltanlou, et al. (2019), Cipora, van Dijck, et al., 2019. Moreover, the confidence intervals for null effects or reliable reversed SNARC effects are typically not wider than for reliable regular SNARC effects, thus pointing at similar measurement precision for reversed, null, and regular effects (see Cipora, Soltanlou, et al., 2019; Cipora, van Dijck, et al., 2019). What is more, we made this observation despite collecting up to 900 data points per experimental cell (30 repetitions per session) for each participant (i.e., a total of 14,400 trials per participant). If this can be confirmed in a larger sample, the original idea of the effects representing some sort of stable mapping of numbers on space (Cipora et al., 2018) needs to be refined. Importantly, this shows that the SNARC and MARC effects should not or at least not purely be considered as trait characteristics at least when measured with the parity judgment task. Considering the SNARC effect as a state instead of a trait can account for the rather mixed pattern of correlations between the SNARC effect and other variables such as mathematics skills, with some studies showing such relationships and others failing to observe them (for an overview, see Table 1 in Cipora, van Dijck, et al., 2020). This might also explain mixed results when it comes to correlations of other cognitive phenomena with other variables (e.g., Hedge et al., 2018).

Taken together, results from correlational studies with these effects can only be generalized when the correlations are replicated between and within samples, and/or situated influences driving the correlations are fully understood. It remains an open question whether other cognitive phenomena are stable across time and whether they are dominant or indominant (Rouder & Haaf, 2018), and we believe our study opens a discussion on this issue.

4.2. Limitations and future research directions

For the interpretation of the study results, several limitations need to be considered. First and most obviously, our sample was relatively small, with the tested individuals not being necessarily representative of the broad population. Unfortunately, the extensive study design with a high time requirement, paired with a commitment to regularly run the experiment, made recruiting and incentivizing participants difficult. At the same time, we believe that the benefits of this approach outweigh the difficulties given the novelty of the insights brought by this method. We wish to encourage researchers to apply the *Ironman paradigm*, because it is in our view the only way to investigate the intraindividual stability of cognitive phenomena. The time requirements and experimental schedule might at first be perceived as a highly inflexible commitment, but in fact this is not the case. Potential participants can be motivated by explaining that the experimental task can be run wherever they have their computer or laptop with them and at whatever time of the day is convenient for them. On top of that, being tired or unmotivated in at least some of the sessions is not perceived as a problem, but instead even leads to the variance in the participant's current state that is desirable from the perspective of the experimenter. This would also reflect the variability between participants current state in typical experimental setups with one session per participant, where some participants are concentrated and motivated, while others are tired, bored, or hungry (e.g., a hangover during participation in a cognitive experimental task due to alcohol consumption on the previous day can affect results, see Murgia et al., 2020).

Second, it must be noted that not all participants were naïve to the research questions of the study; three of the authors of this study also acted as participants. These participants' behavior may be biased, because they knew the research questions and might consciously or unconsciously have had desired or undesired outcomes in their minds while participating. However, we assume that number processing in cognitive speeded RT tasks such as the parity judgment task is highly automatized such that it can hardly be deliberately influenced by expectations or motivation.

Third, participants might have applied different strategies in different sessions, which we have not asked for. The use of different strategies might, however, be related to the size of the effects. Moreover, other covariates than the four that we have tested in the current study might play a role. To explore whether different strategies or covariates explain some of the effects' variability, future studies using the Ironman paradigm could assess task-solving strategies (e.g., whether the participants did "odd" vs. "even" or rather "yes" vs. "no" judgments, focusing only on even; whether they applied some visualization strategies; etc.) or further covariates (e.g., body-related covariates like appetite, context-related covariates like dealing with numbers shortly before the experiment, etc.).

Fourth, it needs to be considered that the mapping of numbers onto space might be a trait rather than a state, but that the SNARC effect

measured in the parity judgment task might not be able to capture this mapping in a stable way. Therefore, we recommend similar studies be undertaken using other tasks in which the SNARC effect can be measured, such as a magnitude classification task.

Lastly, we did not ask about or determine participants' chronotype, which could influence the presence of numerical cognition effects. For instance, a meta-analys by Preckel, Lipnevich, Schneider, and Roberts (2011) found chronotype, in terms of different circadian rhythms and fluctuations in physiological and psychological functions (morningness vs. eveningness), to affect cognitive abilities. Hence, there might be an interaction between chronotype and the covariate of daytime.

Future research should be dedicated to finding reasons for the low intraindividual stability of the SNARC effect in the parity judgment task despite its robustness and high replicability at the group level. The methodology of the present study using the Ironman paradigm combined with bootstrapping approaches should be applied to other cognitive phenomena to answer the following questions: Should other cognitive phenomena be considered as dominant (present in all individuals) or indominant (present at the group level, but absent or even reversed in some individuals), as defined by Rouder and Haaf (2018)? Are other cognitive phenomena intraindividually stable over time and should they be considered as state or trait characteristics of the participants? Which situational contexts are most impactful for intraindividual variability of cognitive phenomena and does this differ between cognitive tasks? Instead of only considering the overall strength of phenomena, as typically done in cognitive psychology, should we also consider their intraindividual variability as an important index, as done in some diagnostic approaches (e.g., attention diagnostics in ADHD) - and if so, when? Are the overall strength of cognitive phenomena (i.e., the mean effect size) correlated with their intraindividual variability?

5. Conclusion

The current study shows that cognitive effects that are robust at the group level might not be present in every individual and are not necessarily stable over time within an individual. This interindividual and intraindividual variability in phenomena like the SNARC effect, the MARC effect, or the Odd Effect should not be considered as error variance or noise. It might not reflect measurement problems either but instead that the underlying phenomena do, in fact, differ between individuals and are not stable over time. Importantly, before using cognitive phenomena as correlate with other variables, their intraindividual stability and situated influences on them must be established. The Ironman paradigm combined with bootstrapping approaches is an innovative and effective study design permitting these relevant and unique insights. Instead of considering cognitive phenomena solely as a trait characteristic of individuals, researchers should think of it in terms of a state characteristic as well.

CRediT authorship contribution statement

Lilly Roth: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Verena Jordan:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Stefania Schwarz:** Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Validation, Writing – original draft, Writing – review & editing. **Klaus Willmes:** Conceptualization, Formal analysis, Validation, Writing – review & editing. **Hans-Christoph Nuerk:** Conceptualization, Project administration, Resources, Validation, Writing – review & editing. **Jean-Philippe van Dijck:** Conceptualization, Resources, Validation, Writing – review & editing. **Krzysztof Cipora:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare no conflicts of interest with the content of this article.

Data availability

All the data, preregistration and the analysis script are shared at https://osf.io/p6u5h/.

Acknowledgements

We would like to thank the participants in this study. We are also grateful to Dror Dotan for his suggestions, as well as to Camilla Gilmore for giving feedback on the earlier version of the manuscript. We would also like to thank Sebastian Sandbrink for proofreading the manuscript.

LR, HCN, and KC are supported by the DFG project "Replicability of Fundamental Results on Spatial-Numerical Associations in Highly Powered Online Experiments (e-SNARC)" (NU 265/8-1). KC is supported by Research England via Centre for Mathematical Cognition.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2024.105781.

Appendix B

Autocorrelations for SNARC and MARC effects

Note that serial correlation coefficients show a systematic bias towards negative values, such that the mean of their distribution is not zero. This bias results from the correlation coefficients and their variance typically not being independently distributed and from using the estimated mean instead of the true mean (Marriott & Pope, 1954). Therefore, as can be seen in our Figs. B1, B2, B3, B4, and C1, CIs for time-series are not symmetrical around zero but shifted towards negative values.

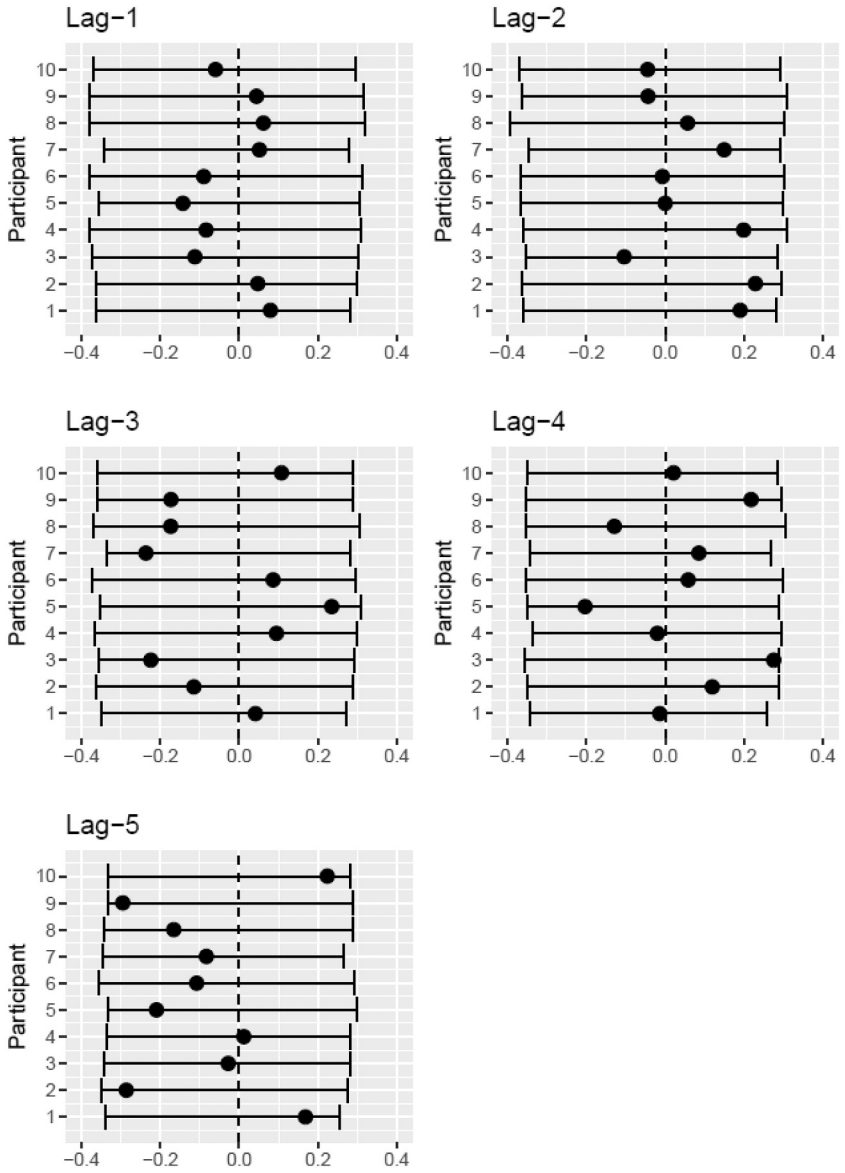


Fig. B1. Autocorrelations and 95% H0 CIs per participant for unstandardized SNARC slopes and different time lags.

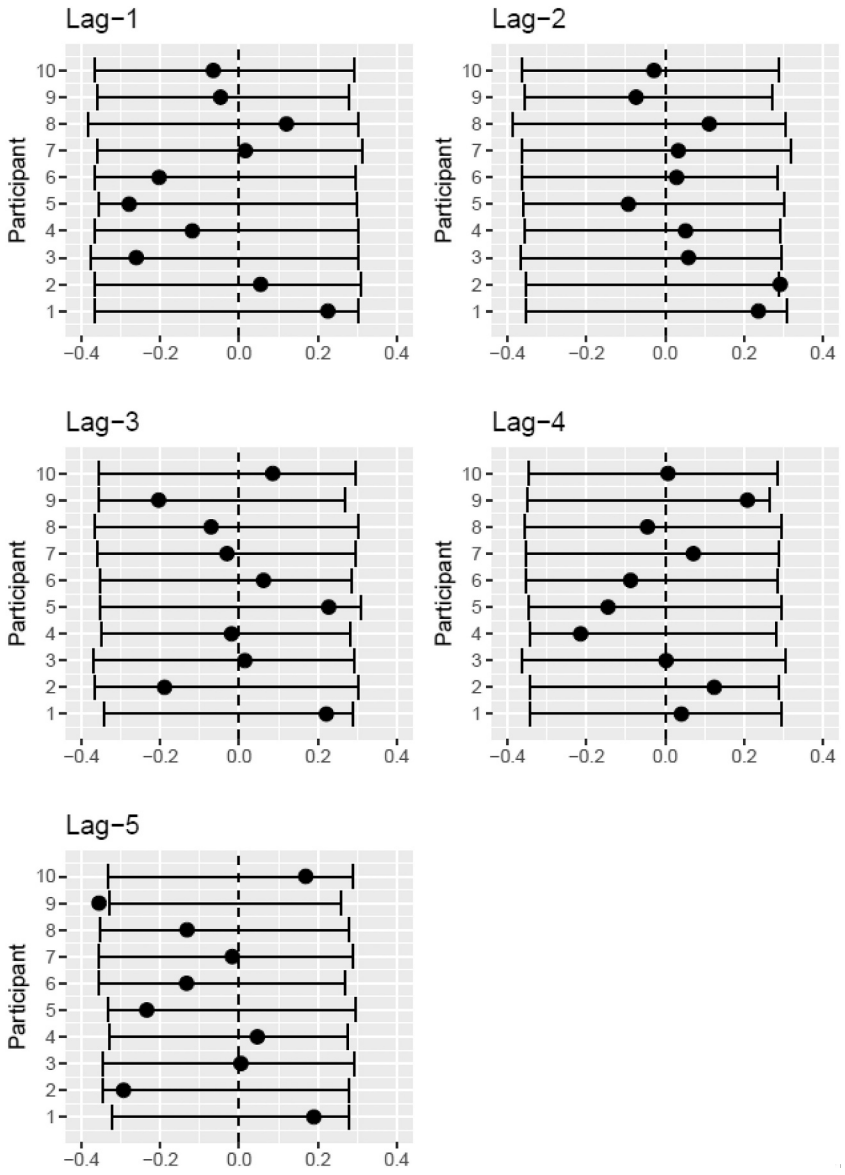


Fig. B2. Autocorrelations and 95% H0 CIs per participant for standardized SNARC slopes and different time lags.

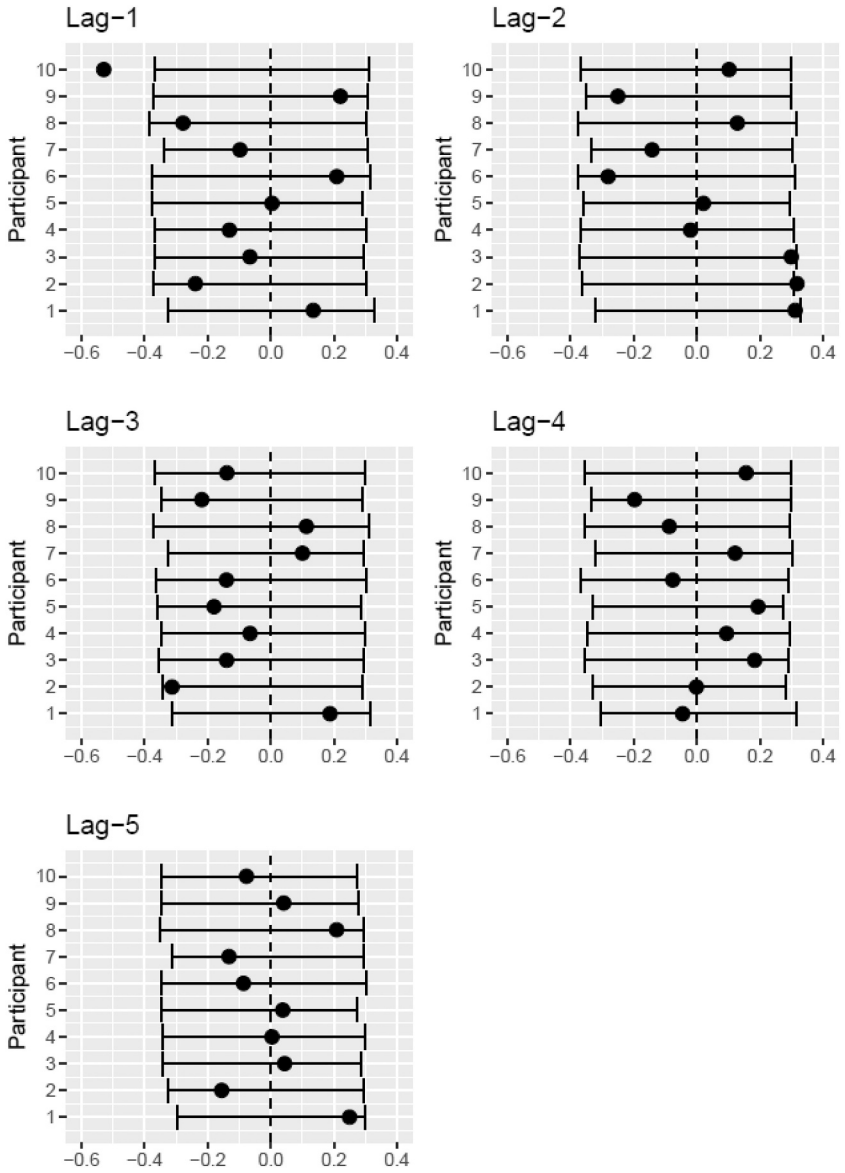


Fig. B3. Autocorrelations and 95% H0 CIs per participant for unstandardized MARC slopes and different time lags.

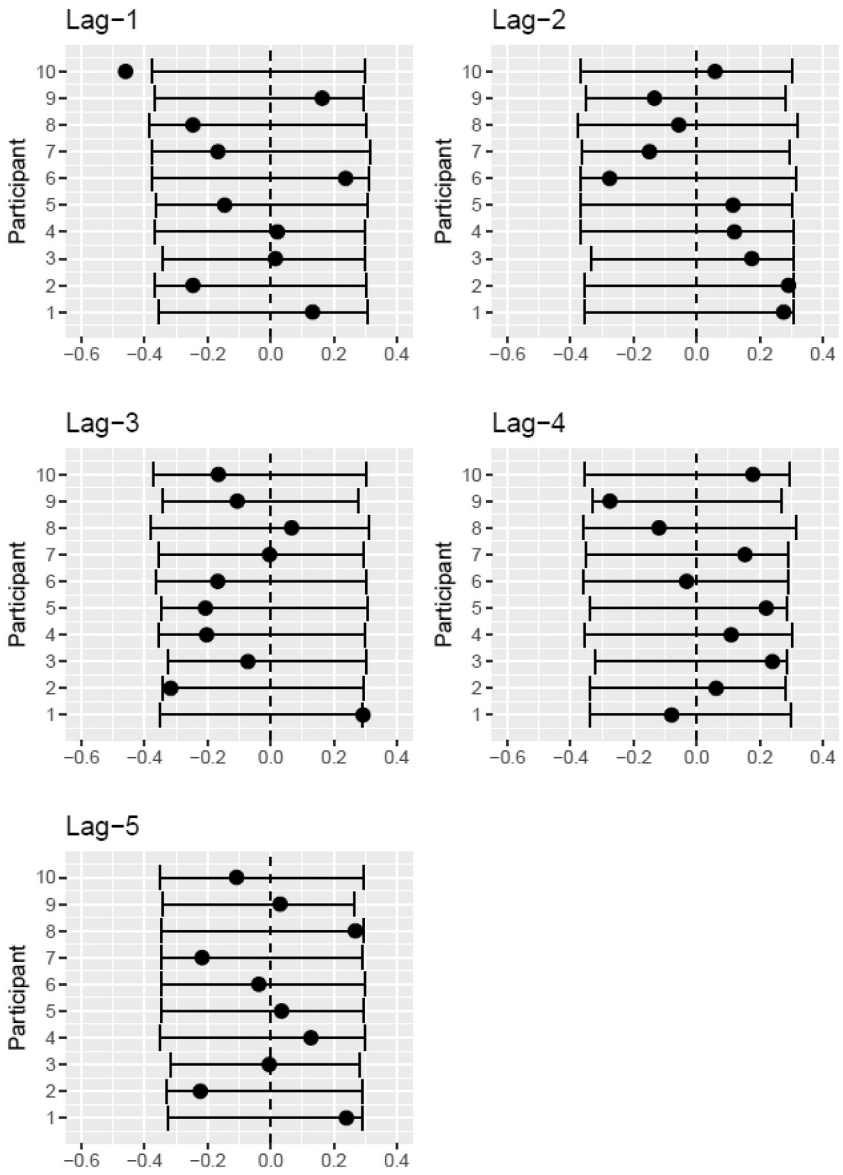


Fig. B4. Autocorrelations and 95% H0 CIs per participant for standardized MARC slopes and different time lags.

Appendix C

C.1. Intraindividual prevalence of the Odd Effect

After seeing the results, we were interested in whether we would find similar intraindividual patterns when analyzing the data regarding different measures or looking into other phenomena one can observe in the parity judgment task. Therefore, we investigated the Odd Effect (i.e., faster responses to even than to odd numbers) obtained by calculating the differences between RTs for odd and for even numbers at the session and the aggregated levels (grand effects) per participant (positive estimates indicating the Odd Effect).

At an aggregated level (grand effects), all ten participants descriptively revealed an Odd Effect (i.e., faster responses to even than to odd numbers). The magnitudes of the grand empirical differences ranged from 0.34 to 18.52 ms. The results were similar for the individual means of session estimates. Without taking CIs into account, the intraindividual proportions of sessions with an Odd Effect ranged from 46.67% to 100.00% (see Table C).

C.2. H0 bootstrapping approach

Subsequently, we conducted H0 bootstrapping analysis to be able to test whether the Odd Effect was more robust than the SNARC and MARC effects. At the 90% confidence level, five out of 10 participants (Participants 1, 3, 5, 7, and 8) showed a reliable Odd Effect at the aggregated level (grand empirical difference), while no participant showed a reliable reversed Odd Effect. The intraindividual proportions of sessions revealing reliable Odd Effects ranged from 10.00% to 86.67%. Two participants showed reliable Odd Effects in more than half of the sessions (Participants 5 and 7). The intraindividual proportions of sessions revealing a reliable reverse Odd Effect ranged from 0.00% to 13.79%. The individual distributions of observed grand empirical and session effects including 90% CIs are visualized in Fig. C.

Table C
Summary of the Odd Effect results. One row corresponds to one participant. The left part shows descriptive properties of the empirical difference in ms (RTs for odd numbers minus RTs for even numbers): proportions of sessions with positive and negative empirical differences (with positive/negative empirical differences reflecting a regular/reversed Odd Effect, respectively); mean empirical difference (averaged from each session), standard deviation, and grand Odd Effect (calculated by aggregating all trials across all sessions). The two following columns correspond to the 90% CI bounds for grand Odd Effects resulting from the H0 bootstrapping approach. The following part summarizes the proportions of sessions with reliable positive, no reliable, and reliable negative Odd Effects.

	Descriptives: Proportions [%]		Descriptives: Empirical differences			90% H0 CIs: Grand		90% H0 CIs: Proportions of sessions [%]		
Participant	Empirical difference > 0	Empirical difference < 0	Mean empirical difference	SD (empirical difference)	Grand empirical difference	lower	upper	Reliable effect > 0	No reliable effect	Reliable effect < 0
1	80.00	20.00	7.57	10.70	7.59	−2.45	2.34	43.33	46.67	10.00
2	56.67	43.33	0.97	6.47	0.89	−1.33	1.35	13.33	83.33	3.33
3	90.00	10.00	7.82	6.36	7.98	−1.72	1.68	33.33	66.67	0.00
4	53.33	46.67	0.12	6.88	0.47	−1.42	1.45	16.67	73.33	10.00
5	93.33	6.67	12.87	9.06	13.32	−1.92	1.85	63.33	36.67	0.00
6	55.17	44.83	2.15	11.06	2.05	−2.27	2.28	20.69	65.52	13.79
7	100.00	0.00	18.56	8.49	18.52	−1.88	1.89	86.67	13.33	0.00
8	78.57	21.43	7.99	11.48	7.31	−3.61	3.58	14.29	82.14	3.57
9[a]	46.67	53.33	0.31	7.16	0.34	−1.52	1.59	10.00	76.67	13.33
10	56.67	43.33	1.50	6.95	1.22	−2.07	2.09	13.33	76.67	10.00

[a] Participant 9 was left-handed.

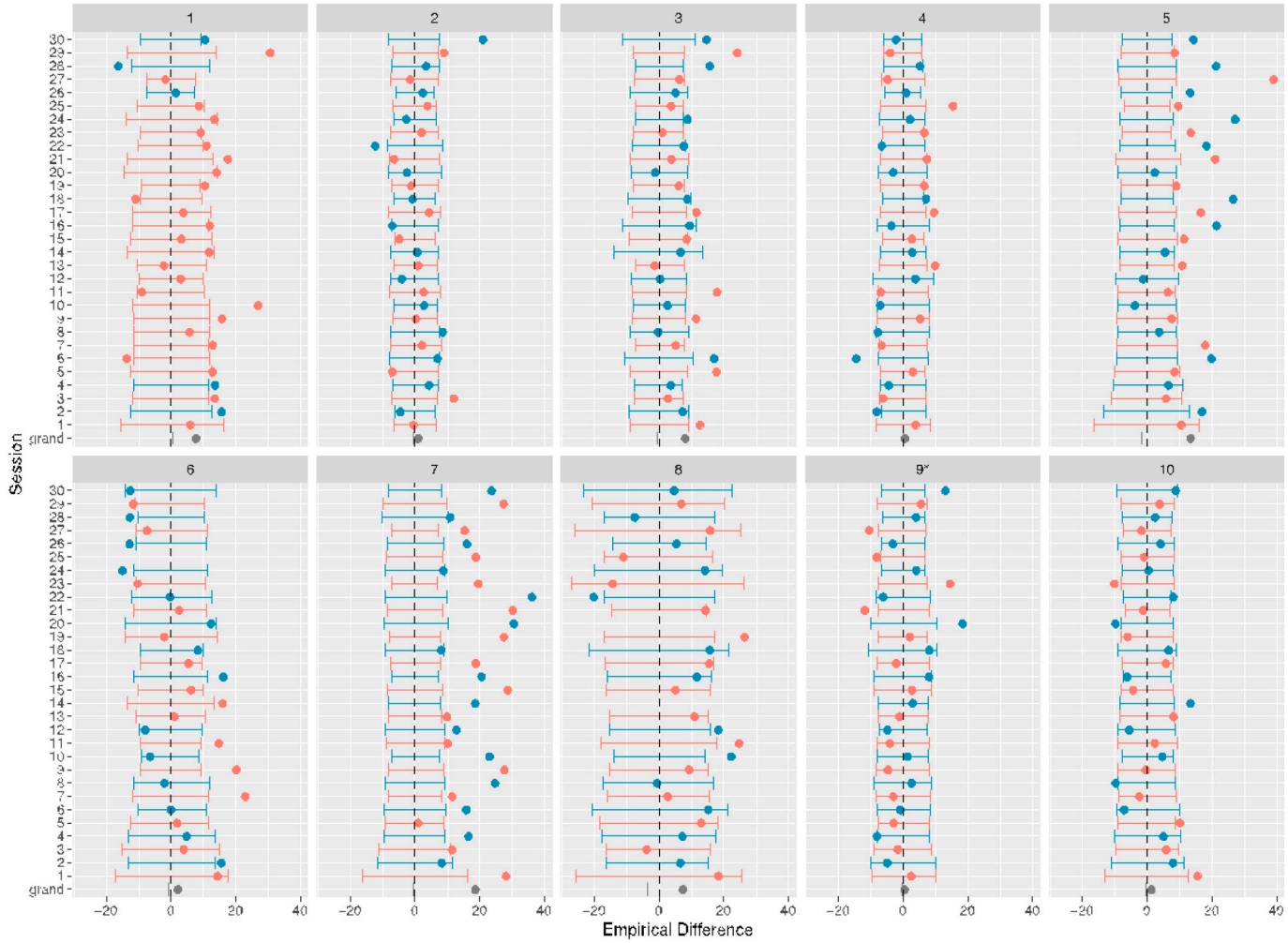


Fig. C. Distributions of observed grand and session Odd Effects (dots) and 90% H0 CIs per participant.
Note. Colors represent the experiment version (red – Version 1: starting with right hand - odd numbers, left hand - even numbers (Block 1), and then reversed; blue – Version 2: reversed order). Note that here only the mean difference between conditions is calculated, so there is no equivalent of standardized slopes in this analysis.
*Participant 9 was left-handed.

Appendix D

D.1. Analyses of mean RTs

Apart from the SNARC, MARC, and Odd Effects, we investigated mean RTs in the parity judgment task. First, we used the H1 bootstrapping approach to calculate session and grand CIs for the mean RTs per participant to analyze RT stability. The individual distributions of mean RTs including 90% H1 CIs at the session and the aggregated levels (grand means) are visualized in Fig. D1. Second, we analyzed the relationship between mean RTs and SNARC/MARC effects by calculating repeated-measures correlations. We observed negative repeated-measures correlation between mean RTs and unstandardized as well as standardized SNARC slopes indicating smaller SNARC effects for shorter RTs. No significant correlations were found between mean RTs and MARC slopes (see Table D1).

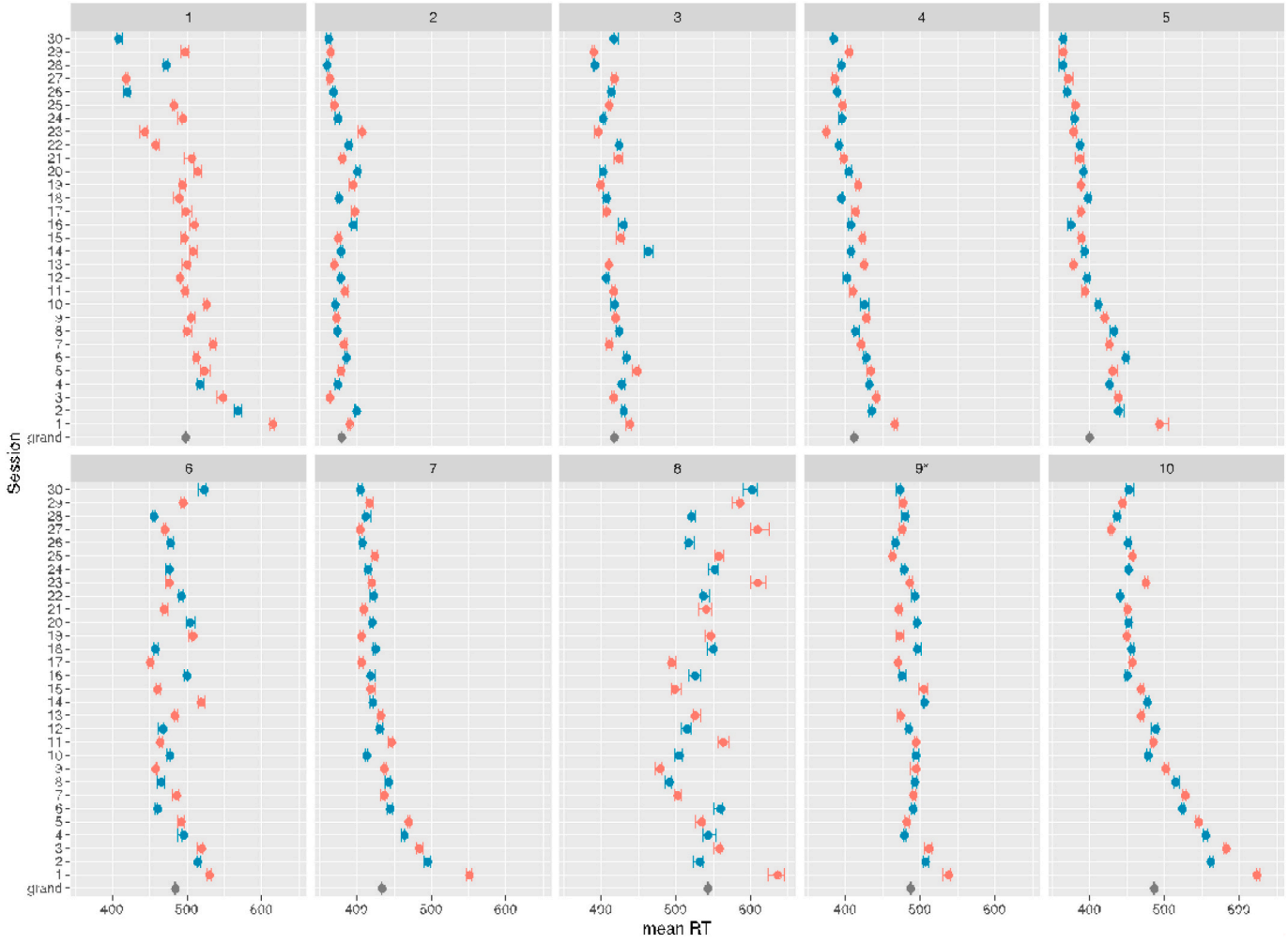


Fig. D1. Distributions of observed grand and session mean RTs (dots) and 90% individual bootstrap CIs per participant. Note that here we only estimated the overall RT per session.
Note. Colors represent the experiment version (red – Version 1: starting with right hand - odd numbers, left hand - even numbers (Block 1), and then reversed; blue – Version 2: reversed order). *Participant 9 was left-handed.

Table D1
Repeated-measures correlations between mean RTs and SNARC/MARC slopes.

Predictors	r	df	p-value	95% CI
Unstandardized SNARC	−0.378	286	<0.001	−0.473, −0.274
Standardized SNARC	−0.270	286	<0.001	−0.318, −0.097
Unstandardized MARC	0.005	286	0.932	−0.111, 0.121
Standardized MARC	0.054	286	0.364	−0.062, 0.168

Lastly, we conducted time-series analysis for mean RTs to analyze autocorrelations. Four out of 10 participants showed a monotone decrease in autocorrelations with an increase in time lags. Furthermore, when looking at Tryon's C statistic, significant positive autocorrelations were revealed by all participants except Participants 6 and 8. When looking at Fig. D1, one can see a clear trend in mean RTs in many participants, such that their responses got faster over the 30 sessions. This systematic trend is in line with the positive autocorrelations and negative repeated-measures

correlations between mean RTs and session number (see Table 5). Table D2 summarizes the autocorrelations between the different observed mean RTs calculated for time series with time lags of one, two, three, four, and five days as well as the corresponding C statistic values.

When looking at the additional H0 bootstrapping analysis of each of the five time-lag autocorrelations, all significant autocorrelations revealed were positive. Three participants (number 4, 5, and 10) showed significant autocorrelations for all time lags. Two more participants showed multiple significant autocorrelations (lag-1, lag-2, lag-3, and lag-4 for Participants 1 and 7). Additionally, Participants 2, 3, and 9 showed a significant lag-1 autocorrelation. Fig. D2 visualizes all autocorrelations and bootstrapped CIs.

Table D2
Autocorrelations in mean RTs for time series with time lags of up to five days between observations (lag-1, lag-2, lag-3, lag-4, lag-5) and Tryon's C statistic.

	Autocorrelations					Tryon's C statistic	
Participant	Lag-1	Lag-2	Lag-3	Lag-4	Lag-5	C	Z
1	0.507	0.300	0.364	0.423	0.205	0.719	3.814*
2	0.395	0.142	0.279	0.086	−0.056	0.441	2.336*
3	0.344	0.015	0.030	0.138	0.154	0.374	1.984*
4	0.602	0.603	0.487	0.490	0.352	0.761	4.036*
5	0.700	0.630	0.521	0.451	0.395	0.887	4.703*
6	0.206	0.079	0.025	−0.108	−0.186	0.328	1.705
7	0.611	0.519	0.386	0.322	0.179	0.857	4.543*
8	0.128	0.282	0.102	0.160	0.062	0.278	1.419
9[a]	0.296	0.199	0.006	0.166	0.121	0.484	2.568*
10	0.778	0.716	0.572	0.458	0.359	0.926	4.909*

[a] Participant 9 was left-handed.
* $p < .025$.

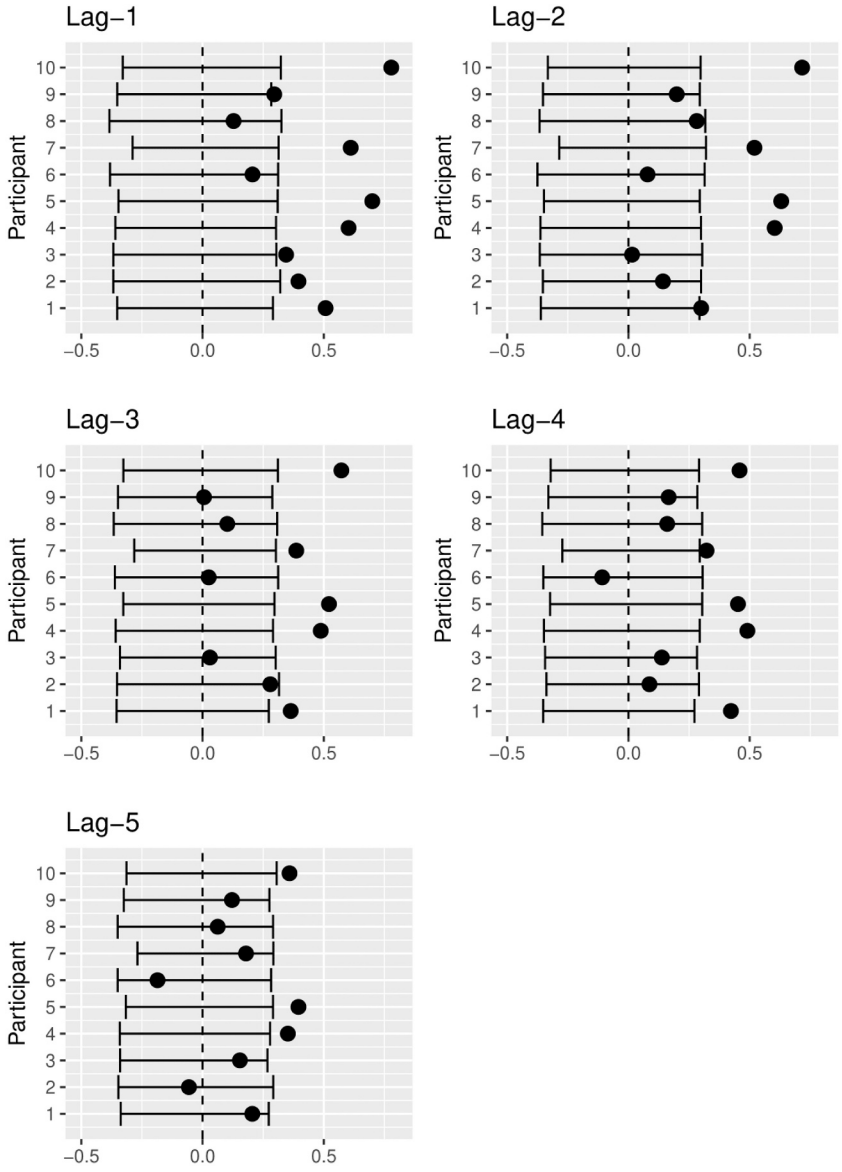


Fig. D2. Autocorrelations and 95% H0 CIs per participant for mean RTs and different time lags.

Appendix E

Sampled group-level t-tests

Each of the sessions in our study could be considered as a single parity judgment experiment as typically reported in the literature. Having seen the large variation of the results across the sessions, we checked how often we could find group-level SNARC, MARC, and Odd effects in our data considering our sample of 10 participants and randomly selected session for each of our participants. We repeated this procedure 5000 times to see what proportion of all 5000 such studies in our sample with only ten participants would have led to significant group-level SNARC, MARC, and Odd effects.

The simulation of 5000 studies with the parity judgment task with one sampled session per participant revealed that at the group level, in one-sided *t*-tests with $\alpha = 0.05$, SNARC and MARC effects could have been found in approximately one out of four cases, while the Odd Effect could have been found in approximately half of the cases (see Table E).

Table E
Proportion of sessions with group-level unstandardized and standardized SNARC and MARC effects and Odd Effect.

Unstandardized SNARC effect	Standardized SNARC effect	Unstandardized MARC effect	Standardized MARC effect	Odd Effect
24.48%	23.96%	23.98%	24.18%	52.12%

Appendix F

Block-order influences on the SNARC and MARC effects

Because two participants had not completed the experiment versions in the correct order, we decided to investigate whether block order influences unstandardized SNARC and MARC effects. More precisely, in Version 1, participants should respond with the right hand for odd and left hand for even numbers in Block 1, and with the right hand for even and left hand for odd numbers in Block 2 (color-coded in red in all figures), while the block order was reversed in Version 2 (color-coded in blue in all figures).

A closer look at Figs. 4, 5, and 6 suggests that a reliable MARC effect was more often observed in Version 2 (color-coded with blue), where participants had to respond with the right hand to even and with the left hand to odd numbers in Block 1 (MARC-compatible), and with the right hand to odd and with the left to even numbers in Block 2 (MARC-incompatible), than in Version 1 (color-coded with red) with the reversed block order. We therefore fit two LMMs for each effect, one of them including only random intercepts per participant (LMM1) and the other including both random intercepts per participant as well as a fixed term for block order (LMM2), which we compared in a likelihood-ratio test (LRT). As visible in Figs. 4, 5, and 6, block order influenced both the unstandardized and standardized MARC effect significantly. The unstandardized MARC slope was on average −6.8 ms in Version 1 (Block 1 MARC-incompatible and Block 2 MARC-compatible), while it was much stronger with −17.0 ms in Version 2 (Block 1 MARC-compatible and Block 2 MARC-incompatible). In contrast, we found no evidence for a block-order influence on the unstandardized and standardized SNARC effect. Note that within each block of the parity judgment task, half of the trials is SNARC-compatible (when odd/even is assigned to the left and the presented number is small) and half of the trials is SNARC-incompatible (when odd/even is assigned to the right and the presented number is large), which provides an explanation for block order not to play a role. Table F displays the exact results of the LRT as well as the parameter estimates for the LMM of choice.

Table F1
Block-order influences (Version 1: MARC-incompatible in Block 1, MARC-compatible in Block 2, Version 2: MARC-compatible in Block 1, MARC-incompatible in Block 2) on unstandardized and standardized SNARC and MARC effects were investigated by comparing two LMMs in a Likelihood-Ratio test (LRT) for each effect: LMM1 including only random intercepts per participant vs. LMM2 including both random intercepts per participant as well as a fixed term for block order. We report both the Akaike information criterion (AIC) and the Bayesian Information Criterion (BIC) for each model as well as the χ^2 statistic (one degree of freedom) and the p-value resulting from the LRT. Moreover, estimates for the model of choice are provided, namely variances of participant intercepts and of residuals as random effects, and slope estimates (with standard errors in brackets) for both versions V1 and V2 together (LMM1) or separately in case of a significant LRT (LMM2) as fixed effects (dummy-coded with V1 as reference).

Model	LRT				Random effects		Fixed effects	
	AIC	BIC	χ^2 (1)	*p*-value	Variance: participants	Variance: residuals	Slope V1 (*SE*)	Slope V2 (*SE*)
Unstandardized SNARC effect								
LMM1	1486.8	1497.9			2.099	7.962	−1.211 (0.487)	
LMM2	1488.7	1503.4	0.156	0.693				
Standardized SNARC effect								
LMM1	313.0	324.1			0.026	0.155	−0.150 (0.056)	
LMM2	314.9	329.7	0.129	0.719				
Unstandardized MARC effect								
LMM1	2948.6	2959.7						
LMM2	2943.8	2958.6	6.768	0.009	134.533	1091.385	−6.826 (4.514)	−10.223 (3.907)

(*continued on next page*)

Table F1 (*continued*)

Model	LRT				Random effects		Fixed effects	
	AIC	BIC	χ^2 (1)	*p*-value	Variance: participants	Variance: residuals	Slope V1 (*SE*)	Slope V2 (*SE*)
Standardized MARC effect								
LMM1	767.0	778.0						
LMM2	760.9	775.7	8.043	0.005	0.084	0.702	−0.168 (0.113)	−0.283 (0.099)

Appendix G

Influence of predictor variables on the SNARC and MARC effects

The influence of predictor variables (sleep, tiredness, stimulants, and daytime) were tested within repeated-measures correlations and LMMs, which included only random participant intercepts, as preregistered. Here, we report LMMs that additionally include random participant slopes for sleep, tiredness, and daytime (see Table G). However, such LMMs were not computed for stimulants, because they would have been very imprecise due to the lack of sessions with prior consumption of stimulants. Importantly, LMMs did not converge for sleep, even when trying our different optimizers, and the variances of random participant slopes were very small for tiredness and daytime and therefore close to singularity. For these reasons, we decided not to compare the fitted LMMs in LRTs and test the influence of the predictor variables for significance, but instead only report all parameter estimates for them in the following table. Significance tests for the models without random participant slopes can be found in the repeated-measures correlations and LMMs in the main part of the paper

Table G

On top of repeated-measures correlations or LMMs with random participant intercepts, the influence of sleep duration, current level of tiredness, and time of the day on unstandardized and standardized SNARC and MARC effects was also investigated in LMMs that additionally included random participant slopes. While sleep duration and current level of tiredness were linear variables, time of the day was categorically coded with morning (M), afternoon (A), and evening (E), with M dummy-coded as reference. We report both the Akaike information criterion (AIC) and the Bayesian Information Criterion (BIC) for each model. Moreover, our table displays estimates for variances of participant intercepts and participant slopes and of residuals as random effects, and slope estimates (with standard errors in brackets) as fixed effects.

	Model fit		Random effects			Fixed effects
Predictor	AIC	BIC	Variance: Participant intercepts	Variance: Participant slopes	Variance: Residuals	Slope estimates (SE)
Unstandardized SNARC effect						
Sleep	1485.6	1507.8	0.310	0.068	7.886	0.005 (0.185)
Tiredness	1488.1	1510.3	4.806	0.252	7.484	−0.195 (0.187)
Daytime	1490.5	1527.5	M: 1.120	A: 1.179	7.646	M: −0.902 (0.431)
				E: 0.069		A: −0.724 (0.527)
						E: −0.344 (0.407)
Standardized SNARC effect						
Sleep	314.0	336.1	0.046	0.002	0.152	<0.001 (0.027)
Tiredness	318.2	340.3	0.018	<0.001	0.155	−0.009 (0.013)
Daytime	316.1	353.1	M: 0.013	A: 0.024	0.148	M: −0.121 (0.052)
				E: <0.001		A: −0.089 (0.074)
						E: −0.016 (0.056)
Unstandardized MARC effect						
Sleep	2941.2	2963.3	2315.214	36.667	1082.253	−1.866 (2.776)
Tiredness	2953.9	2976.1	101.663	0.105	1114.983	−0.716 (1.055)
Daytime	2959.4	2996.3	M: 132.020	A: 11.671	1102.688	M: −11.282 (4.906)
				E: 16.670		A: −3.351 (4.882)
						E: 2.858 (5.038)
Standardized MARC effect						
Sleep	766.3	788.4	<0.001	0.001	0.711	−0.081 (0.051)
Tiredness	771.4	793.6	0.088	<0.001	0.717	−0.033 (0.027)
Daytime	776.3	813.3	M: 0.073	A: 0.062	0.700	M: −0.234 (0.120)
				E: 0.041		A: −0.146 (0.143)
						E: −0.040 (0.139)

References

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology, 8*, 1–13. https://doi.org/10.3389/fpsyg.2017.00456

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods, 26*(3), 295–314. https://doi.org/10.1037/met0000337

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Belli, F., & Fischer, M. H. (2023). Breathing shifts visuo-spatial attention. *Cognition, 243*. https://doi.org/10.1016/j.cognition.2023.105685. Article 105685.

Bland, J. M., & Altman, D. G. (2015). Statistics notes: Bootstrap resampling methods. *BMJ, 350*, Article h2622. https://doi.org/10.1136/bmj.h2622

Cevada, T., Conde, E., Marques, D., & Deslandes, A. C. (2019). Test-retest reliability of the simon task: A short version proposal. *Somatosensory & Motor Research, 36*(4), 275–282. https://doi.org/10.1080/08990220.2019.1689114

Cipora, K., & Göbel, S. M. (2013). Number-space associations: Just how reliable is the SNARC effect [poster presentation]. In *16th European Workshop on Cognitive Neuropsychology*.
Cipora, K., He, Y., & Nuerk, H.-C. (2020). The spatial–numerical association of response codes effect and math skills: Why related? *Annals of the New York Academy of Sciences, 1477*(1), 5–19. https://doi.org/10.1111/nyas.14355
Cipora, K., Hohol, M., Nuerk, H.-C., Willmes, K., Brożek, B., Kucharzyk, B., & Nęcka, E. (2016). Professional mathematicians differ from controls in their spatial-numerical associations. *Psychological Research, 80*, 710–726. https://doi.org/10.1007/s00426-015-0677-6
Cipora, K., & Nuerk, H.-C. (2013). Is the SNARC effect related to the level of mathematics? No systematic relationship observed despite more power, more repetitions, and more direct assessment of arithmetic skill. *Quarterly Journal of Experimental Psychology, 66*(10), 1974–1991. https://doi.org/10.1080/17470218.2013.772
Cipora, K., Patro, K., & Nuerk, H.-C. (2015). Are spatial-numerical associations a cornerstone for arithmetic learning? The lack of genuine correlations suggests no. *Mind, Brain, and Education, 9*(4), 190–206. https://doi.org/10.1111/mbe.12093
Cipora, K., Patro, K., & Nuerk, H.-C. (2018). Situated influences on spatial–numerical associations. In T. Hubbard (Ed.), *Spatial biases in perception and cognition* (pp. 41–49). Cambridge: Cambridge University Press.
Cipora, K., Soltanlou, M., Reips, U.-D., & Nuerk, H.-C. (2019). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behavior Research Methods, 51*(4), 1676–1692. https://doi.org/10.3758/s13428-019-01213-5
Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Göbel, S. M., Willmes, K., … Nuerk, H.-C. (2019). A minority pulls the sample mean: On the individual prevalence of robust group-level cognitive phenomena – The instance of the SNARC effect. *PsyArXiv*. https://doi.org/10.31234/osf.io/bwyr3
Cipora, K., van Dijck, J.-P., Georges, C., Masson, N., Schiltz, C., Pesenti, M., Göbel, S. M., Willmes, K., & Nuerk, H.-C. (2020). Parity and space: On the prevalence and individual differences in the MARC effect. In *XXXVIIIth European Workshop on Cognitive Neuropsychology, Bressanone, Italy*. https://www.osf.io/4qhpw.
Cipora, K., & Wood, G. (2017). Finding the SNARC instead of hunting it: A 20* 20 Monte Carlo investigation. *Frontiers in Psychology, 8*, 1194. https://doi.org/10.3389/fpsyg.2017.01194
Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671–684. https://doi.org/10.1037/h0043943
Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*(2), 116–127. https://doi.org/10.1037/h0076829
Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The neuronal recycling hypothesis. In S. Dehaene, J. R. Duhamel, M. D. Hauser, & G. Rizzolatti (Eds.), *From monkey brain to human brain* (pp. 133–157). MIT Press.
Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*(3), 371–396. https://doi.org/10.1037/0096-3445.122.3.371
Didino, D., Breil, C., & Knops, A. (2019). The influence of semantic processing and response latency on the SNARC effect. *Acta Psychologica, 196*, 75–86. https://doi.org/10.1016/j.actpsy.2019.04.008
Eide, P., Kemp, A., Silberstein, R. B., Nathan, P. J., & Stough, C. (2002). Test-retest reliability of the emotional stroop task: Examining the paradox of measurement change. *The Journal of Psychology, 136*(5), 514–520. https://doi.org/10.1080/00223980209605547
Endler, N. S., & Kocovski, N. L. (2001). State and trait anxiety revisited. *Journal of Anxiety Disorders, 15*(3), 231–245. https://doi.org/10.1016/S0887-6185(01)00060-3
Fias, W., Brysbaert, M., Geypens, F., & D'Ydewalle, G. (1996). The importance of magnitude information in numerical processing: Evidence from the SNARC effect. *Mathematical Cognition, 2*(1), 95–110. https://doi.org/10.1080/135467996387552
Fischer, M. H., Mills, R. A., & Shaki, S. (2010). How to cook a SNARC: Number placement in text rapidly changes spatial-numerical associations. *Brain and Cognition, 72*(3), 333–336. https://doi.org/10.1016/j.bandc.2009.10.010
Fischer, M. H., & Shaki, S. (2014). Spatial associations in numerical cognition-from single digits to arithmetic. *Quarterly Journal of Experimental Psychology, 67*(8), 1461–1483. https://doi.org/10.1080/17470218.2014.927515
Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences, 115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115
Georges, C., Hoffmann, D., & Schiltz, C. (2013). The SNARC effect and its relationship to spatial abilities in women [poster presentation]. In *18th conference of the European Society for Cognitive Psychology, Budapest, Hungary*. http://hdl.handle.net/10993/13012.
Gevers, W., Santens, S., Dhooge, E., Chen, Q., van den Bossche, L., Fias, W., & Verguts, T. (2010). Verbal-spatial and visuospatial coding of number-space interactions. *Journal of Experimental Psychology: General, 139*(1), 180–190. https://doi.org/10.1037/a0017688
Gevers, W., Verguts, T., Reynvoet, B., Caessens, B., & Fias, W. (2006). Numbers and space: A computational model of the SNARC effect. *Journal of Experimental Psychology: Human Perception and Performance, 32*(1), 32–44. https://doi.org/10.1037/0096-1523.32.1.32
Guida, A., Mosinski, F., Cipora, K., Mathy, F., & Noël, Y. (2020). Spatialization in working memory: Can individuals reverse the cultural direction of their thoughts? *Annals of the New York Academy of Sciences, 1477*(1), 113–125. https://doi.org/10.1111/nyas.14499
Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review, 26*(3), 772–789. https://doi.org/10.3758/s13423-018-1522-x
Happé, F., & Frith, U. (2006). The weak coherence account: Detail-focused cognitive style in autism Spectrum disorders. *Journal of Autism and Developmental Disorders, 36*(1), 5–25. https://doi.org/10.1007/s10803-005-0039-0
Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1
Hines, T. M. (1990). An odd effect: Lengthened reaction times for judgments about odd digits. *Memory and Cognition, 18*(1), 40–46. https://doi.org/10.3758/BF03202644
Hohol, M., Willmes, K., Nęcka, E., Brożek, B., Nuerk, H.-C., & Cipora, K. (2020). Professional mathematicians do not differ from others in the symbolic numerical distance and size effects. *Scientific Reports, 10*(1), 11531. https://doi.org/10.1038/s41598-020-68202-z
Hommel, B., Fischer, R., Colzato, L. S., van den Wildenberg, W. P. M., & Cellini, C. (2012). The effect of fMRI (noise) on cognitive control. *Journal of Experimental Psychology: Human Perception and Performance, 38*(2), 290–301. https://doi.org/10.1037/a0026353
Jensen, A. R. (2006). *Clocking the mind. Mental chronometry and individual differences*. Elsevier Ltd.. https://doi.org/10.1016/B978-0-08-044939-5.X5000-9
Judd, N., Aristodemou, M., & Kievit, R. (2023). Cognitive variability is ubiquitous and distinct from mean performance across eleven tasks with over 7 million trials. *Open Science Framework*. https://doi.org/10.31234/osf.io/b29rn. Preprint.
Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13
Lindemann, O., Abolafia, J. M., Pratt, J., & Bekkering, H. (2008). Coding strategies in number space: Memory requirements influence spatial-numerical associations. *The Quarterly Journal of Experimental Psychology, 61*(4), 515–524. https://doi.org/10.1080/17470210701728677
Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 149–157. https://doi.org/10.1037/0278-7393.16.1.149
Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Reading: Addison-Wesley.
Luck, S. (2019). *Why experimentalists should ignore reliability and focus on precision*. Luck Lab. Retrieved from https://lucklab.ucdavis.edu/blog/2019/2/19/reliability-and-precision.
Lyons, I. M., Nuerk, H.-C., & Ansari, D. (2015). Rethinking the implications of numerical ratio effects for understanding the development of representational precision and numerical processing across formats. *Journal of Experimental Psychology: General, 144*(5), 1021–1035. https://doi.org/10.1037/xge0000094
Marriott, F. H. C., & Pope, J. A. (1954). Bias in the estimation of autocorrelations. *Biometrika, 41*(3/4), 390–402. https://doi.org/10.2307/2332719
Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7
McManus, R. M., Young, L., & Sweetman, J. (2023). Psychology is a property of persons, not averages or distributions: Confronting the group-to-person generalizability problem in experimental psychology. *Advances in Methods and Practices in Psychological Science, 6*(3). https://doi.org/10.1177/25152459231186615. Article 25152459231186615.
Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin Review, 20*, 819–858. https://doi.org/10.3758/s13423-013-0404-5
Murgia, M., Mingolo, S., Prpic, V., Sors, F., Santoro, I., Bilotta, E., & Agostini, T. (2020). University Students' hangover may affect cognitive research. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.573291. Article 573291.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*(3), 353–383. https://doi.org/10.1016/0010-0285(77)90012-3
Ninaus, M., Moeller, K., Kaufmann, L., Fischer, M. H., Nuerk, H.-C., & Wood, G. (2017). Cognitive mechanisms underlying directional and non-directional spatial-numerical associations across the lifespan. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.01421. article 1421.
Nuerk, H.-C., Iversen, W., & Willmes, K. (2004). Notational modulation of the SNARC and the MARC (linguistic markedness of response codes) effect. *Quarterly Journal of Experimental Psychology, 57A*, 835–863. https://doi.org/10.1080/02724980343000512
Paap, K. R., & Sawi, O. (2016). The role of test-retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods, 274*, 81–93. https://doi.org/10.1016/j.jneumeth.2016.10.002
Paap, K. R., Wagner, S., Johnson, H., Bockelman, M., Cushing, D., & Sawi, O. (2014). *20,000 flanker trials: Are the effects reliable, robust, and stable? [poster]*. San Francisco, USA: Association for Psychological Science.
Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science, 2*(4), 378–395. https://doi.org/10.1177/2515245919879695
Poirel, N., Pineau, A., Jobard, G., & Mellet, E. (2008). Seeing the Forest before the trees depends on individual field-dependency characteristics. *Experimental Psychology, 55*(5), 328–333. https://doi.org/10.1027/1618-3169.55.5.328

Preckel, F., Lipnevich, A. A., Schneider, S., & Roberts, R. D. (2011). Chronotype, cognitive abilities, and academic achievement: A meta-analytic investigation. *Learning and Individual Differences, 21*(5), 483–492. https://doi.org/10.1016/j.lindif.2011.07.003

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science, 1*(1), 19–26. https://doi.org/10.1177/2515245917745058

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review, 26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Schaefer, J., Opgen-Rhein, R., & Strimmer, K. (2021). *GeneNet: Modeling and inferring gene networks. R package version 1.2.16*. https://CRAN.R-project.org/package=GeneNet.

Schuch, S., Philipp, A. M., Maulitz, L., & Koch, I. (2022). On the reliability of behavioral measures of cognitive control: Retest reliability of task-inhibition effect, task-preparation effect, Stroop-like interference, and conflict adaptation effect. *Psychological Research, 86*(7), 2158–2184. https://doi.org/10.1007/s00426-021-01627-x

Shaki, S., & Fischer, M. H. (2008). Reading space into numbers: A cross-linguistic comparison of the SNARC effect. *Cognition, 108*(2), 590–599. https://doi.org/10.1016/j.cognition.2008.04.001

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Skinner, B. F. (1966). Operant behavior. In W. K. Honig (Ed.), *Operant behavior: Areas of research and application* (pp. 12–32). New York: Appleton-Century-Crofts.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review, 25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test-retest reliability of standard and emotional stroop tasks: An investigation of color-word and picture-word versions. *Assessment, 12*(3), 330–337. https://doi.org/10.1177/1073191105276375

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651

Toomarian, E. Y., & Hubbard, E. M. (2018). On the genesis of spatial-numerical associations: Evolutionary and cultural factors co-construct the mental number line. *Neuroscience & Biobehavioral Reviews, 90*, 184–199. https://doi.org/10.1016/j.neubiorev.2018.04.010

Tryon, W. W. (1982). A simplified time-series analysis for evaluating treatment interventions. *Journal of Applied Behavior Analysis, 15*(3), 423–429. https://doi.org/10.1901/jaba.1982.15-423

Van Dijck, J.-P., Abrahamse, E., & Fias, W. (2020). Do preliterate children spontaneously employ spatial coding for serial order in working memory? *Annals of the New York Academy of Sciences, 1477*(1), 91–99. https://doi.org/10.1111/nyas.14430

Van Dijck, J.-P., Fias, W., & Cipora, K. (2022). Spatialization in working memory and its relation to math anxiety. *Annals of the New York Academy of Sciences, 1512*(1), 192–202. https://doi.org/10.1111/nyas.14765

Viarouge, A., Hubbard, E. M., & McCandliss, B. D. (2014). The cognitive mechanisms of the SNARC effect: An individual differences approach. *PLoS One, 9*(4), Article e95756. https://doi.org/10.1371/journal.pone.0095756

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software, 40*(1), 1–29. https://doi.org/10.18637/jss.v040.i01

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. https://ggplot2.tidyverse.org.

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation. R package version 1.1.0*. https://CRAN.R-project.org/package=dplyr.

Wickham, H., Vaughan, D., & Girlich, M. (2023). *Tidyr: Tidy messy data. R package version 1.3.0*. https://CRAN.R-project.org/package=tidyr.

Wilke, C. (2020). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.1.1*. https://CRAN.R-project.org/package=cowplot.

Wilson, R. E., Thompson, R. J., & Vazire, S. (2017). Are fluctuations in personality states more than fluctuations in affect? *Journal of Research in Personality, 69*, 110–123. https://doi.org/10.1016/j.jrp.2016.06.006

Wood, G., Nuerk, H. C., & Willmes, K. (2006b). Variability of the SNARC effect: Systematic interindividual differences or just random error? *Cortex, 42*(8), 1119–1123. https://doi.org/10.1016/S0010-9452(08)70223-5

Wood, G., Nuerk, H.-C., & Willmes, K. (2006a). Crossed hands and the SNARC effect: A failure to replicate Dehaene, Bossini and Giraux (1993). *Cortex, 42*(8), 1069–1079. https://doi.org/10.1016/S0010-9452(08)70219-3

Wood, G., Willmes, K., Nuerk, H.-C., & Fischer, M. H. (2008). On the cognitive link between space and number: A meta-analysis of the SNARC effect. *Psychology Science Quarterly, 50*(4), 489–525.

World Medical Association. (2013). Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA, 310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences, 45*, E1. https://doi.org/10.1017/S0140525X20001685

Zayas, V., Sridharan, V., Lee, R. T., & Shoda, Y. (2019). Addressing two blind spots of commonly used experimental designs: The highly-repeated within-person approach. *Social and Personality Psychology Compass, 13*(9), Article e12487. https://doi.org/10.1111/spc3.12487

Zorowitz, S., & Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging, 8*(8), 789–797. https://doi.org/10.1016/j.bpsc.2023.02.004

A7. A computational contextualization model

Citation:

Roth, L.*, Hassan, G.*, Cipora, K., Nuerk. H.-C., & Butz, M. V. (in preparation). *A computational contextualization model for the SNARC and SNARC-like effects.*

* shared first authorship

We used the Python programming language, version 3.12.9 (Python Software Foundation, 2025), for coding the currently presented version of our computational SNARC model, as well as the following libraries:

- Matplotlib, version 3.10.1 (Hunter, 2007)
- NumPy, version 2.2.3 (Harris et al., 2020)
- Pandas, version 2.2.3 (The pandas development team, 2020)
- Scipy, version 1.15.2 (Virtanen et al., 2020)

A7.1. Details on spatial bias distributions

In this part, the spatial bias distributions in C0, C1, C2, and C4 for number tasks (see Figure 3-2) and in C2 and C3 for letter tasks (see Figure 3-3) are described in greater detail than in Section 3.3.1. To start with, we decided to model the spatial bias distributions using normal distributions because we assumed symmetric variance (characterized by the SD) around the expected bias (characterized by the mean). The normal distributions were truncated due to the restriction to the interval from −1 (leftmost cognitive spatial bias) to 1 (rightmost cognitive spatial bias). As the area under each probability distribution should equal 1, this truncation leads to different maximal probability densities between distributions. The means and SD for each distribution in each schema are hard-coded in the current version of our computational model. Because cognitive context schemata with spatial bias distributions are theoretical constructs and thus not directly observable, we inserted values that we considered plausible; we may optimize these values in the future.

In C0, C2, and C3, each spatial bias distribution reflected one stimulus, whereas in C1 and C4, some distributions were set to hold for several stimuli. C0 includes eight spatial bias distributions for the typically used stimulus set from 1 to 9 (excluding 5), which we used in all three number tasks (CJ, MC, and PJ) in our model. The distance between distribution means is proportional to the stimulus magnitude, and they are symmetrically located around 0. C1 comprises a distribution one for number magnitudes smaller than 5 and another distribution for those larger than 5, with their means being equally spaced from 0. C2 is the only cognitive context schema that we postulate for both number and letter stimuli. It includes spatial bias distributions for ordinal positions in WM, which can either refer to ordinal stimulus positions within a newly learned sequence stored in WM, or within the currently used stimulus set. The distance between distribution means is proportional to the ordinal position, and they are symmetrically located around 0. In the three number tasks we have implemented in the model, C2 includes eight equally spaced distributions for the 1st (i.e., number 1) to 8th (i.e., number 9) ordinal position within the stimulus set, making C2 look similar to C0. In the three letter tasks, C2 comprises only four equally spaced distributions. In CVC with low WM salience, they correspond to the 1st (e.g., letter A in the set A, J, O, R) to 4th (e.g., letter R in the set A, J, O, R) ordinal position within the stimulus set (see Experiment 1 in Study 4). In CVC with intermediate or high WM salience, they reflect the 1st (e.g., letter R in the sequence RAOJ) to 4th (e.g., letter J in the sequence RAOJ)

ordinal position within the newly memorized sequence instead (see Experiments 2 and 3 in Study 4). Spatial bias distributions were asymmetrically located in C3 and C4. In C3, we implemented a strong left bias distribution for letter A. We let distances between distribution means decrease for subsequent letters in the alphabet up until letter G, keep the distances equal for letters G to T, and increase them again up until letter Z. In C4, separate distributions reflecting left biases were defined for numbers 1, 2, and 3, whereas all numbers from 4 to 9 were summarized in one distribution reflecting a slight right bias. These four distributions were equally spaced. For an illustration of spatial bias distributions, please refer to Figures 3-2 (C0, C1, C2, and C4 for number tasks) and 3-3 (C2 and C3 for letter tasks).

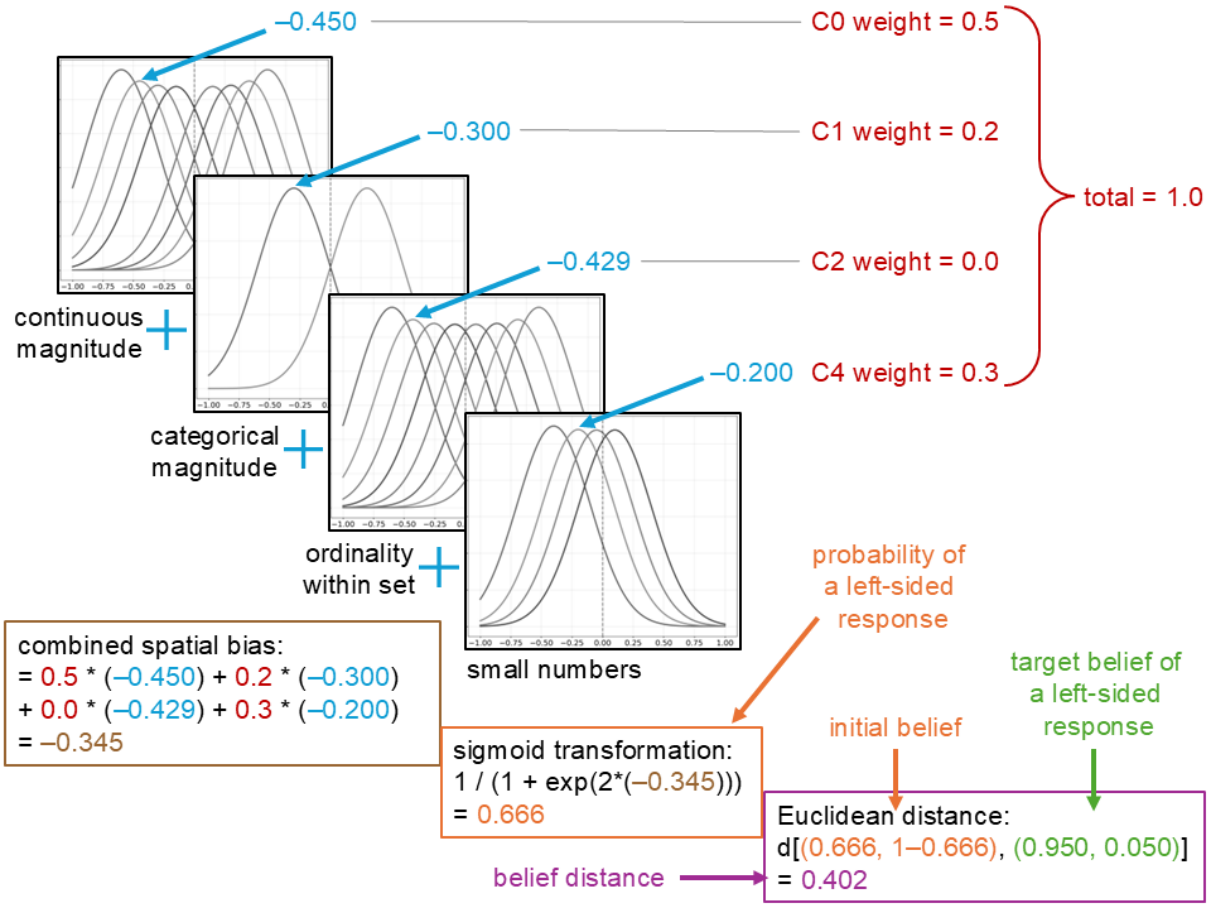
A7.2. Details on data simulation

Our Python code simulates, preprocesses, and stores trial-level data. It additionally analyzes and plots sample-level SNARC and SNARC-like effects. In this section, I will provide details on the exact functionalities and underlying algorithms. Users may first choose a specific task, sample size, and number of repetitions per experimental cell (e.g., MC with 200 participants and 25 repetitions). Alternatively, they may use the default settings, which initiates a run of all six tasks for the sample sizes and number of repetitions that we used in Studies 1, 2, and 4 in this dissertation. Specifically, the default sample sizes are 349 in nominal CJ, 662 in MC and PJ, and 162, 109, and 117 in CVC with low, intermediate, and high WM salience, respectively (i.e., sample sizes remaining after participant-level exclusions based on criteria such as age or seriousness in participation, but before trial-level exclusions due to incorrect, too fast, or too slow responses). Default numbers of repetitions are 30 in all number tasks, and 20 in all letter tasks.

To simulate data for specific tasks, the respective stimuli (i.e., numbers or letters) and their attributes (i.e., blue or yellow in nominal CJ, go or no-go in CVC tasks) are assigned to simulated participants upon task selection. Similarly, further experimental manipulations are set up (i.e., two blocks differing in response-to-key assignments, counterbalanced block orders across participants, and a letter sequence in CVC with intermediate or high WM salience). Additionally, all weights for the cognitive contexts that can be activated (i.e., C0, C1, C2, and C4 in number tasks; C2 and C3 in letter tasks) are initialized with values greater than 0. The simulation of RT and given responses is explained in the following, and important computational steps are illustrated for an example trial in Figure A7-1.

Figure A7-1

Belief distance computation for stimulus number 2 in an example trial



Note. This figure was created in PowerPoint (Microsoft Corporation, 2024). The schematic illustration of a belief distance computation holds for an example trial in any number task, where the number 2 is presented, and a left-sided response is required according to the task instructions. For simplicity reasons, the mean of each spatial bias distribution (blue) is used for this example RT calculation, although the computational contextualization model can also account for variations when simulating data. The upper part of the figure shows how spatial biases from several cognitive context schemata are weighted (red) and summed up to a combined spatial bias (brown), as done in the box in the lower left part. Note that only C0, C1, C2, and C4 are considered, because C3 is specific to letter stimuli. The combined spatial bias is then sigmoid transformed (using a scaling factor of 2), which can be seen in the box in the lower centered part, and the resulting value reflects the probability of a left-sided response (orange). Together with the opposite probability, namely that of a right-sided response, it constitutes the initial belief (orange). Finally, the belief distance (purple) is computed as the Euclidean distance between the initial belief and the target belief of a correct left-sided response (green).

During the simulation process, the first step in each trial is the computation of the combined spatial bias as the weighted sum of biases for the respective stimulus across all activated cognitive context schemata. The combined bias lies within the range from −1 (maximal left bias) to 1 (maximal right bias), consistent with the spatial

biases within each schema. In the next step, it is sigmoid transformed (using a scaling factor of 2, instead of the most commonly used scaling factor of −1) to yield a value between 0 and 1, which can be interpreted as the probability for a left-sided response. The two-dimensional vector of the response probabilities (left, right) will be referred to as the "initial belief" henceforth. At the same time, we assume that the threshold that must be reached to trigger a correct response is a probability of 0.95. Hence, we define another two-dimensional vector that we call "target belief," which is (0.95, 0.05) if a left-sided response is required and (0.05, 0.95) if a right-sided response is required as per the task instructions. The RT and error probability for the trial are then calculated based on the Euclidean distance between the initial belief and the target belief, referred to as "belief distance." In the current version of our model, this relationship is linear on average: The RT is computed as the sum of the baseline RT and the belief distance, which depends on the context weights, as it is influenced by the initial belief, which itself is determined by those context weights. Moreover, the belief distance is multiplied by a "sensitivity factor," scaling the extent to which the initial belief interferes with the target belief. Finally, normally distributed noise is added according to the SD of RT:

$$RT_{ijk} \sim N\left(RT_{baseline\ ij} + sensitivity_{conflict\ ij} * distance_{belief\ ijk},\ SD^2_{RT\ ij}\right)$$

with i = participant, j = task, k = stimulus

The error probability is proportional to the belief distance, and a response is sampled based on the resulting error probability for the respective trial. Note that we have to date not yet optimized a baseline error probability, a scaling factor for sensitivity to cognitive conflict, or added any random noise. That is, out of all optimized parameters, so far, only context weights play a role for the error probability and thus for the given response in each trial. The logic underlying this simulation mechanism is that the further the initial belief is from the target belief (i.e., the larger the belief distance), the slower the response and the higher the error probability. The Python implementation automatically uses optimized parameters if available; otherwise, it builds the computations onto default parameters. The trial-by-trial dataset is automatically stored as a text file.

A7.3. Details on parameter optimization

Different sets of parameters can be optimized for the computational contextualization model for the SNARC and SNARC-like effects. Parameters under question are the context weights, the baseline RT, the SD of RT, and the sensitivity to

cognitive conflict. As I will explain in the following, the suitability of several of these parameters for optimization may be subject to discussion.

The most important goal of the parameter optimization for this computational context model is to shed light on the extents to which each cognitive context schema we postulate is activated in different tasks. To figure out their relative influence on behavioral patterns, we operationalize these extents of activation as weights that sum up to 1 within each task. For example, let us assume that the estimated weights for C0 and C4 were both 0.5 in CJ (implying that all other context weights were 0). This would suggest that the SNARC effect in CJ is determined by two cognitive context schemata to an equal extent. Specifically, this would indicate that the SNARC effect in CJ is half driven by spatial biases that continuously range from left to right, proportionally aligning with number magnitude, and that it is additionally half driven by leftward biases for small numbers. In this example, the categorization into numbers smaller or larger than 5 as reflected by C1 and the ordinality of each number within the stimulus set as reflected by C2 would appear to be irrelevant. Optimizing these context weights lies at the core of our computational contextualization model: Knowing the relative contributions of the different schemata can provide crucial insights into cognitive mechanisms underlying the SNARC and SNARC-like effects.

Moreover, optimizing the baseline RT seemed meaningful and unproblematic to us, because this is necessary to account for general response speed differences between individuals. Additionally, one can optimize the SD of RT (i.e., the individual response speed variability), and the sensitivity to cognitive conflict. From a theoretical and psychological point of view, it might seem necessary to optimize these two parameters as well, because large interindividual differences exist in human response speed variability and sensitivity to cognitive conflict behavioral patterns. Yet, from a practical and mathematical point of view, optimizing these two additional parameters might be problematic regarding parameter identifiability. That is, the optimization procedure may fail to identify the optimal parameter combinations of the baseline RT, the SD of RT, and the sensitivity to cognitive conflict in a stable way, because of trade-offs between the parameters that enable different combinations to lead to similar fits. Notably, optimizations of baseline RT truly turned out more robust when the SD of RT and the sensitivity factor were not simultaneously optimized (for details, see Tables A7-5 and A7-6 in Appendix A7.4.).

To conclude, while optimizing context weights and the baseline RT at the individual level seemed unproblematic or even crucial to us, both advantages and disadvantages exist to optimizing the SD of RT and the sensitivity to cognitive conflict. We therefore completed two runs that differed in the set of parameters to be optimized: one run targeting the context weights, the baseline RT, the SD of RT, and the sensitivity to cognitive conflict (referred to as “Optimization 1” in the following), and one run targeting only the context weights and the baseline RT (referred to as “Optimization 2” in the following). Preliminary results reported in the Section 3.3.2. were obtained with Optimization 1, whereas this Appendix also includes results obtained with Optimization 2.

All parameters were optimized at the individual level for each participant separately using preprocessed trial-level data from Studies 1 (CJ), 2 (MC and PJ), and 4 (CVC with low, intermediate, and high WM salience). That is, the current version of our Python code first removes no-go trials, incorrectly answered trials, and those yielding RT outliers (below 200 ms or above 1500 ms as fixed exclusion criteria, and ±3 individual SD beyond the individual mean in an iterative procedure for each participant). Note that this preprocessing pipeline was similar but not equal to the ones used in each of our studies, but our aim was to keep it consistent across the six tasks for the current purpose. Each optimization run consisted of two steps: First, we ran a global stochastic optimization algorithm (i.e., differential evolution from the SciPy library; Virtanen et al., 2020) to explore the entire parameter space and avoid getting trapped in local minima. Second, we ran a local gradient-based optimization algorithm (i.e., L-BFGS-B from the SciPy library; Virtanen et al., 2020) to fine-tune the optimization and more precisely find the minimum. We used the root mean squared error (RMSE) as an objective function and compared the baseline RMSE using default parameters with the RMSE after optimization for each participant (see Table A7-2) to ensure that parameters were improved. Default parameters reflect theoretical assumptions made based on behavioral patterns observed in Studies 1, 2, and 4, and can be considered rather exploratory; they are displayed in Table A7-1.

Table A7-1

Default parameters for each task in the computational contextualization model

Task	M context weights (SD)	Baseline RT	SD of RT	Sensitivity factor
Nominal CJ	C0: 0.300 (0), C1: 0.001 (0), C2: 0.001 (0), C3: 0 (0), C4: 0.698 (0)	450 ms	100 ms	30
MC	C0: 0.001 (0), C1: 0.798 (0), C2: 0.001 (0), C3: 0 (0), C4: 0.200 (0)	570 ms	125 ms	30
PJ	C0: 0.698 (0), C1: 0.001 (0), C2: 0.001 (0), C3: 0 (0), C4: 0.300 (0)	650 ms	125 ms	30
CVC with low WM salience	C0: 0 (0), C1: 0 (0), C2: 0.001 (0), C3: 0.999 (0), C4: 0 (0)	550 ms	150 ms	30
CVC with intermediate WM salience	C0: 0 (0), C1: 0 (0), C2: 0.100 (0), C3: 0.900 (0), C4: 0 (0)	530 ms	150 ms	30
CVC with high WM salience	C0: 0 (0), C1: 0 (0), C2: 0.500 (0), C3: 0.500 (0), C4: 0 (0)	580 ms	150 ms	30

Note. The context weights refer to the relative influence of the respective cognitive context schema onto the response pattern. The baseline RT and SD RT refer to the offset and variability of RT. The sensitivity factor is a measure for the extent to which the belief distance influences RT and error probabilities. The same default parameters were used for the simulation of all trials for all participants, as reflected by an SD of 0 for all context weights. In number tasks, the weight for C3 was fixed at 0, and in letter tasks, the weights for C0, C1, and C4 were fixed at 0.

A7.4. Detailed preliminary optimization results

This section includes observations regarding the optimization goodness (i.e., RMSE improvement report), optimized parameters (i.e., average parameters across participants and their variability), and a robustness test (i.e., comparison between multiple optimization runs) both optimization approaches.

The average RMSE across participants for default parameters were substantially higher than the average RMSE for optimized parameters in each task (for exact RMSE means, see Table A7-2), indicating that the optimization was fruitful. Specifically, both optimization algorithms were able to reduce the average RMSE across participants by 50.3% to 72.1% in each task. Surprisingly, although Optimization 1 included more parameters (i.e., the SD of RT and the sensitivity to cognitive conflict) than Optimization 2, the RMSE improvements were similar with both procedures. This may indicate that the parameters common to both procedures (i.e., the context weights and baseline RT) were mainly responsible for the improvement. Moreover, similar RMSE improvements suggest that the simpler algorithm, Optimization 2, can be used instead of the more complex one, Optimization 1, in further developments of this project. Importantly, although both algorithms seemed to work well, this does not prove that the model captures truly underlying cognitive mechanisms: Wrongly assumed cognitive context schemata or wrongly defined cognitive spatial bias distributions may be able to produce realistic behavioral patterns. We consider Optimizations 1 and 2 as promising first attempts, but we may investigate further cognitive context schemata or additionally optimize the spatial bias distribution parameters they contain in the future.

Table A7-2

Mean RMSE across participants per task for default versus optimized parameters in the computational contextualization model

Task	RMSE with default parameters	RMSE after Optimization 1	RMSE after Optimization 2
Nominal CJ	67.58 ms	20.26 ms	18.85 ms
MC	110.85 ms	33.95 ms	35.36 ms
PJ	128.59 ms	37.16 ms	38.49 ms
CVC (low)	77.23 ms	33.35 ms	34.60 ms
CVC (interm.)	73.69 ms	32.22 ms	32.32 ms
CVC (high)	91.11 ms	44.81 ms	45.32 ms

Note. The CVC descriptions in parentheses in the leftmost column refer to low, intermediate, and high WM salience in these tasks using letter stimuli.

Averaged parameters across participants resulting from Optimizations 1 (context weights as included in Section 3.3.2., baseline RT, SD of RT, and sensitivity to cognitive conflict) and 2 (context weights and baseline RT) are displayed in Tables A7-3 and A7-4, respectively. Importantly, optimized context weights and baseline RT were similar between both algorithms, suggesting that the optimization of these parameters is mostly independent from the optimization of the SD of RT and the sensitivity factor. Interestingly, Optimization 1 lead to slightly different estimates for the sensitivity factor than the default ones, which were used for Optimization 2. The preliminary results suggest that the sensitivity to cognitive conflict is higher in CJ compared to MC and PJ, and in CVC with WM load compared to CVC without WM load. Notably the robustness of sensitivity factors across optimization runs seemed poor in the cases where sensitivity factors appeared high (see Table A7-5).

Table A7-3

Parameters per task averaged across participants resulting from Optimization 1 of the computational contextualization model

Task	M context weights (SD)	Baseline RT	SD RT	Sensitivity factor
Nominal CJ	C0: 0.240 (0.123), C1: 0.273 (0.129), C2: 0.245 (0.120), C4: 0.241 (0.126)	404.22 ms	152.91 ms	46.57
MC	C0: 0.042 (0.123), C1: 0.573 (0.455), C2: 0.125 (0.288), C4: 0.260 (0.413)	514.37 ms	151.33 ms	22.96
PJ	C0: 0.026 (0.087), C1: 0.216 (0.367), C2: 0.070 (0.194), C4: 0.688 (0.412)	562.28 ms	152.15 ms	27.94
CVC with low WM salience	C2: 0.247 (0.410), C3: 0.753 (0.410)	561.94 ms	149.99 ms	23.90
CVC with intermediate WM salience	C2: 0.836 (0.327), C3: 0.164 (0.327)	537.36 ms	143.26 ms	40.02
CVC with high WM salience	C2: 0.868 (0.310), C3: 0.132 (0.310)	585.63 ms	158.44 ms	42.87

Note. For the purpose of summarizing general tendencies, this table displays the means (for all parameters) and SD (for context weights) across participants, although they were optimized at the individual level. The context weights, the baseline RT, the SD of RT, and the sensitivity to cognitive conflict were targeted in Optimization 1. Context weights refer to the relative influence of the respective cognitive context schema onto the response pattern. Context weights were fixed at 0 for C3 in number tasks and for C0, C1, and C4 in letter tasks, which is why they are not displayed in this table. The baseline RT and SD RT refer to the offset and variability of RT. The sensitivity factor is a measure for the extent to which the belief distance influences RT and error probabilities. Note that the optimized context weights are identical with those reported in Section 3.3.2.

Table A7-4

Parameters per task averaged across participants resulting from Optimization 2 of the computational contextualization model

Task	M context weights (SD)	Baseline RT	SD RT	Sensitivity factor
Nominal CJ	C0: 0.273 (0.130), C1: 0.251 (0.132), C2: 0.257 (0.120), C4: 0.219 (0.119)	414.70 ms	100.00 ms (default)	30 (default)
MC	C0: 0.058 (0.194), C1: 0.514 (0.448), C2: 0.084 (0.235), C4: 0.343 (0.433)	509.04 ms	125 ms (default)	30 (default)
PJ	C0: 0.040 (0.147), C1: 0.168 (0.314), C2: 0.061 (0.190), C4: 0.731 (0.388)	559.58 ms	125 ms (default)	30 (default)
CVC with low WM salience	C2: 0.256 (0.403), C3: 0.744 (0.403)	556.91 ms	150 ms (default)	30 (default)
CVC with intermediate WM salience	C2: 0.800 (0.376), C3: 0.200 (0.376)	544.03 ms	150 ms (default)	30 (default)
CVC with high WM salience	C2: 0.835 (0.355), C3: 0.165 (0.355)	593.52 ms	150 ms (default)	30 (default)

Note. For the purpose of summarizing general tendencies, this table displays the means (for all parameters) and SD (for context weights) across participants, although they were optimized at the individual level. Only the context weights and the baseline RT were targeted in Optimization 2, while defaults were kept for the SD of RT and the sensitivity to cognitive conflict, which is why the values are equal to those displayed in Table A7-1. Context weights refer to the relative influence of the respective cognitive context schema onto the response pattern. Context weights were fixed at 0 for C3 in number tasks and for C0, C1, and C4 in letter tasks, which is why they are not displayed in this table. The baseline RT and SD RT refer to the offset and variability of RT. The sensitivity factor is a measure for the extent to which the belief distance influences RT and error probabilities.

To assess whether the optimization results are stable, we ran a robustness test on Optimizations 1 and 2. The test consisted in optimizing parameters 10 times while setting different random seeds, creating 10 sets of “digital twins” (i.e., sets of optimized parameters generated 10 times under the same conditions). In each run, all parameters were optimized for 10 randomly drawn participants per tasks. The variability between digital twins within each participant serves as a measure of robustness. A higher parameter variability indicates that the optimizer finds different solutions across runs. A lower parameter variability indicates that the optimizer converges to similar parameters across runs, and we were striving for such stability. Beyond optimized parameters, we also assessed the stability of the RMSE between runs. A higher RMSE variability indicates that the optimizer sometimes finds low-quality solutions, converging to local rather than global minima. A lower RMSE variability indicates that optimizer finds high-quality solutions in most runs, converging to global minima. Tables A7-5 and A7-6 display the variability of each optimized parameter and resulting RMSE obtained in the robustness tests for Optimizations 1 and 2, respectively, averaged across the 10 randomly drawn participants in each task.

Table A7-5

Variability of parameters and RMSE averaged across participants in Optimization 1 in the computational contextualization model

Task	Context weights	Baseline RT	SD RT	Sensitivity factor	RMSE
Nominal CJ	C0: 0.132, C1: 0.113, C2: 0.119, C4: 0.127	9.15 ms	51.99 ms	13.06	2.89 ms
MC	C0: 0.041, C1: 0.003, C2: 0.075, C4: 0.045	1.31 ms	48.96 ms	1.79	0.01 ms
PJ	C0: 0.018, C1: 0.054, C2: 0.025, C4: 0.055	3.16 ms	49.64 ms	4.72	0.01 ms
CVC (low)	C2 and C3: 0.019	1.00 ms	53.57 ms	1.25	0.01 ms
CVC (interm.)	C2 and C3: 0.005	12.06 ms	49.35 ms	18.87	<0.01 ms
CVC (high)	C2 and C3: 0	12.39 ms	49.84 ms	19.45	0.01 ms

Note. Parameter optimization was run at an individual level. As a measure of robustness, the SD of parameter estimates and RMSE across optimization runs was therefore computed within participants. This table, however, displays the means of these SD (i.e., the variability within participants averaged across participants). The CVC descriptions in parentheses in the leftmost column refer to low, intermediate, and high WM salience in these tasks using letter stimuli.

Table A7-6

Variability of parameters and RMSE averaged across participants in Optimization 2 in the computational contextualization model

Task	Context weights	Baseline RT	RMSE
Nominal CJ	C0: 0.126, C1: 0.127, C2: 0.124, C4: 0.100	2.39 ms	1.35 ms
MC	C0: 0.030, C1: 0.017, C2: 0.019, C4: 0.004	0.21 ms	< 0.01 ms
PJ	C0: 0.012, C1: 0.019, C2: 0.016, C4: 0.044	0.25 ms	0.01 ms
CVC (low)	C2 and C3: 0	0.38 ms	<0.01 ms
CVC (interm.)	C2 and C3: 0	0.33 ms	<0.01 ms
CVC (high)	C2 and C3: <0.001	0.44 ms	<0.01 ms

Note. Parameter optimization was run at an individual level. As a measure of robustness, the SD of parameter estimates and RMSE across optimization runs was therefore computed within participants. This table, however, displays the means of these SD (i.e., the variability within participants averaged across participants). The CVC descriptions in parentheses in the leftmost column refer to low, intermediate, and high WM salience in these tasks using letter stimuli.

The robustness test revealed that optimized context weights were relatively stable across runs in both Optimizations 1 and 2. While this was the case for five of the tasks, stability seemed lower for CJ. Importantly, the absence of variability in most instances of the CVC tasks points towards a flaw in the current implementation. When taking a closer look at individual optimization results in each run, one can see that this zero variability results from either of the two cognitive context schemata being weighted

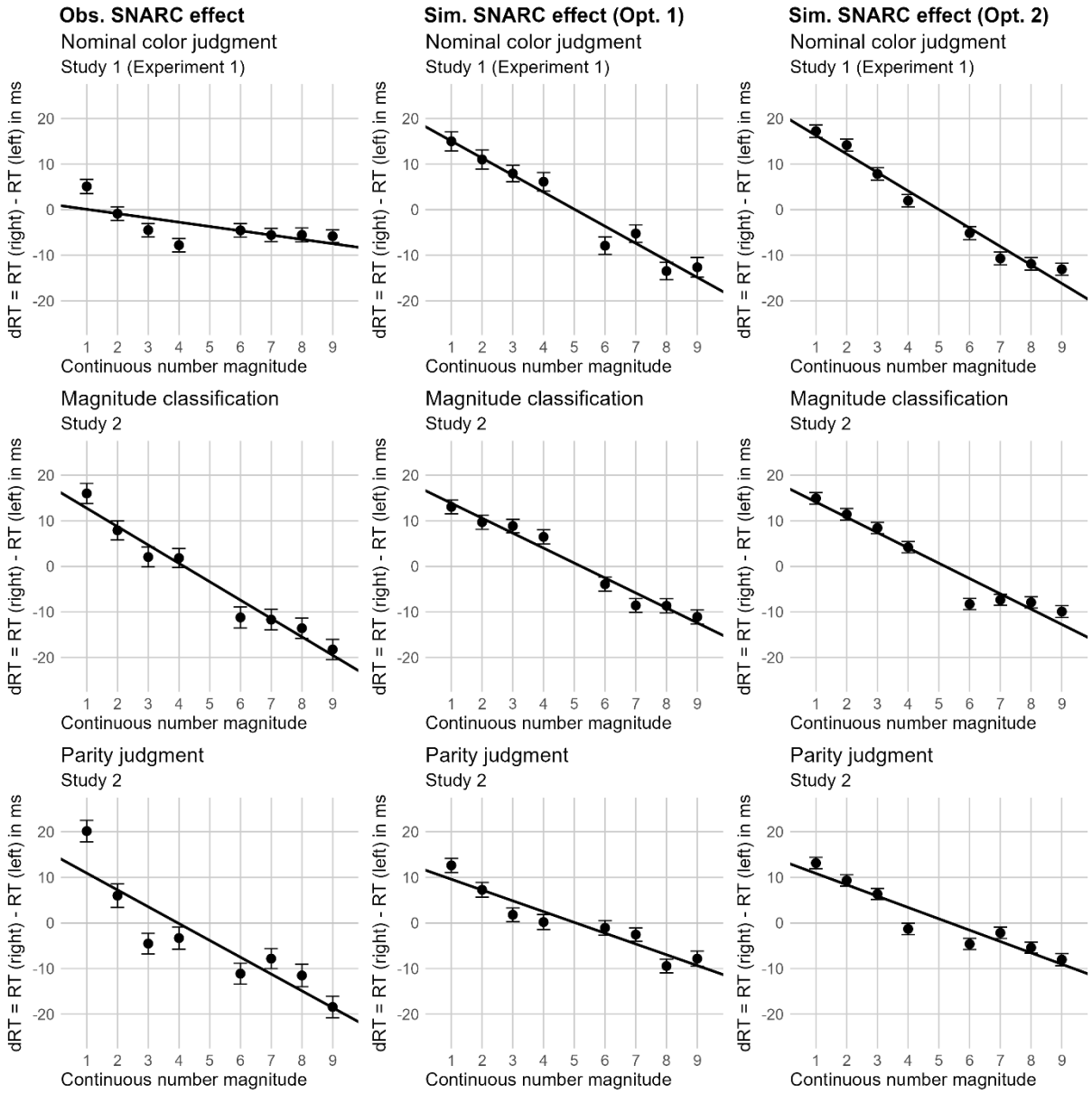
with 0, and the other with 1. This may, on the one hand, be caused by the randomized assignment of participants to go stimulus sets and memorized sequences by the model, ignoring the real experimental setup and the assignment in previous runs. That is, while we aimed at creating 10 sets of digital twins, there was mismatch of experimental setups between optimization runs for most participants, which may undermine the results. On the other hand, extreme context weights hitting the boundaries of 0 or 1 were also obtained in number tasks, where all participants were assigned to the same stimulus set. We will look into this issue and explore alternative optimization algorithms in the future to reveal combinations of simultaneously activated schemata instead. Notably, the RMSE variability observed in the robustness test was very low, suggesting that the optimization procedure converges towards the same solutions across runs. While this is not surprising for letter tasks due to the previously described mismatch, the high parameter robustness in number tasks is promising.

A7.5. Comparison of observed data and simulated data

As described in Section 3.3.2., for each task, we visualized simulated data obtained with both preliminary optimization attempts next to the empirically observed data, so that dRT patterns can be compared. Visualizations for the SNARC effect in three number tasks, the LTM-based OPE in three letter tasks, and the WM-based OPE in two letter tasks can be found in Figures A7-2, A7-3, and A7-4. The middle and right columns in the figures show that the computational contextualization model successfully produces the effects in each task. Note that dRT patterns in the plots in the left column, showing empirically observed data, may look slightly different from those illustrated in our published manuscript. The reason for this is that slightly different preprocessing pipelines were used across studies, whereas we consistently used the same preprocessing pipeline across all six tasks for the parameter optimization.

Figure A7-2

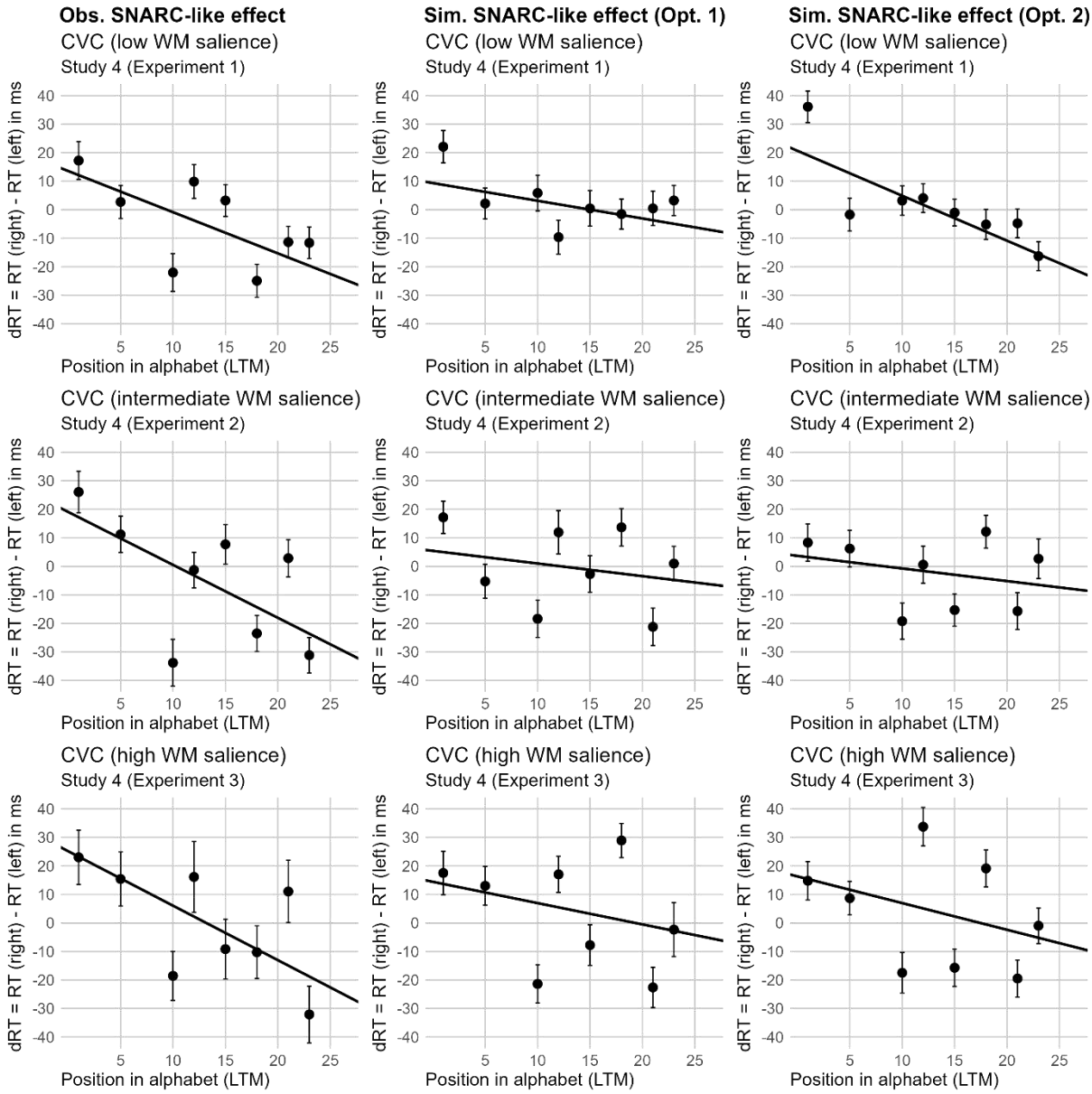
Comparison of observed and simulated SNARC effects in three number tasks



Note. This figure was created in R (R Core Team, 2025). It shows nine plots, each of which is a typical illustration of the SNARC effect in tasks using numbers as stimuli. Therein, dRT averaged across all participants (with bars reflecting standard errors of the mean dRT) on the y-axis are shown as a linear function of continuous number magnitude on the x-axis. Plots in the left column show the empirically observed data, whereas plots in the middle and right columns show the simulated data predicted by the computational contextualization model using parameters that resulted from Optimizations 1 and 2, respectively. Plots in the upper, middle, and lower row illustrate data from nominal CJ, MC, and PJ, respectively. Note that the left and middle column show the exact same plots as Figure 3-4 does in Section 3.3.2.

Figure A7-3

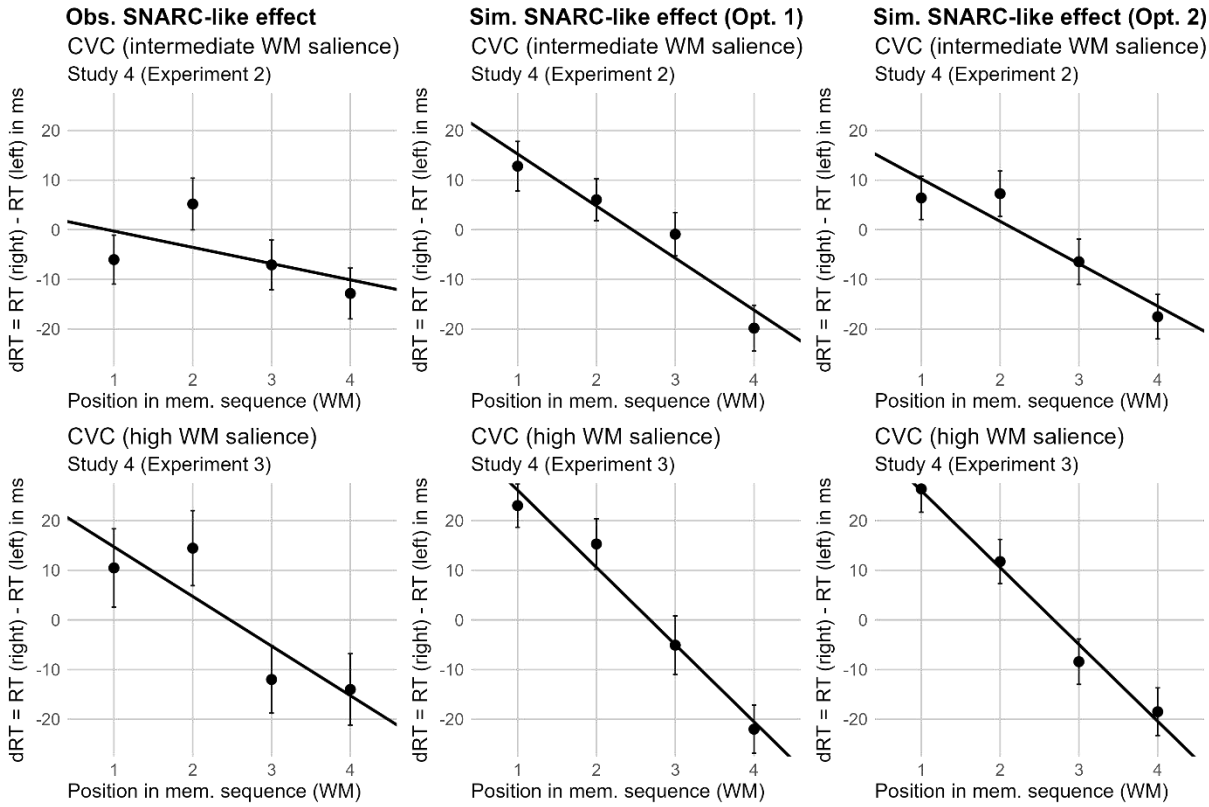
Comparison of observed and simulated LTM-based OPE in three letter tasks



Note. This figure was created in R (R Core Team, 2025). It shows nine plots, each of which is a typical illustration of the LTM-based OPE in tasks using letters as stimuli. Therein, dRT averaged across all participants (with bars reflecting standard errors of the mean dRT) on the y-axis are shown as a linear function of letter positions in the alphabet on the x-axis. Note that each participant was only assessed with four letters, but all data from all eight stimuli is summarized within the displayed regression lines for simplicity. Plots in the left column show the empirically observed data, whereas plots in the middle and right columns show the simulated data predicted by the computational contextualization model using parameters that resulted from Optimizations 1 and 2, respectively. Plots in the upper, middle, and lower row illustrate data from CVC with low, intermediate, and high WM salience, respectively.

Figure A7-4

Comparison of observed and simulated WM-based OPE in two letter tasks



Note. This figure was created in R (R Core Team, 2025). It shows six plots, each of which is a typical illustration of the WM-based OPE in tasks using letters as stimuli. Therein, dRT averaged across all participants (with bars reflecting standard errors of the mean dRT) on the y-axis are shown as a linear function of letter positions in the memorized sequence on the x-axis. Note that participants were assessed with different letter sequences, but all data is summarized within the plots including the displayed regression lines for simplicity. Plots in the left column show the empirically observed data, whereas plots in the middle and right columns show the simulated data predicted by the computational contextualization model using parameters that resulted from Optimizations 1 and 2, respectively. Plots in the upper and lower row illustrate data from CVC with intermediate and high WM salience, respectively.

Finally, an additional test was conducted to ensure that the data simulated by the computational contextualization model using individually optimized parameters resembles the empirically observed data. For this, I calculated correlations between the simulated data and the corresponding empirically observed data for each participant. As a baseline correlation for comparison, I randomly split the empirically observed data into two halves and calculated the correlation between randomly matched participants. The rationale is that, if the simulation produces meaningful datasets, the correlations between the simulated data and the corresponding

empirically observed data should exceed the correlation between empirically observed data for randomly matched participants. All correlations were computed across all six tasks, using data at the aggregation level of dRT per stimulus and per participant (i.e., eight data points per participant in number tasks, and four data points per participant in letter tasks). With this aggregation level, the impact of noise is reduced compared to raw data, while preserving data structure in terms of the experimental cells for each participant.

Importantly, these calculations were computed for data simulated in the "deterministic mode" of our Python code, in which two sources of variability are eliminated. The RT in each trial is computed by using the mean of each spatial distribution (i.e., where probability density is maximal) and the average RT prediction (i.e., the sum of the baseline RT and the belief distance weighted by the sensitivity factor without adding noise based on the SD of RT). This was meaningful for the computation of the correlations to avoid searching for patterns in data that is intentionally made noisy. In contrast, in the "variable mode," random variability based on SD in spatial bias distributions and task-specific SD of RT is switched on. Data simulated in the variable mode may better reflect natural randomness in real experimental data, thereby increasing external validity. The simulated data illustrated in Figures 3-4, A7-2, A7-3, and A7-4 were simulated using the variable mode.

The correlations between simulated data and the corresponding empirically observed data were $r = 0.158$ for Optimization 1 and $r = 0.159$ for Optimization 2. In contrast, the correlation between empirically observed data for randomly matched participants, which served as a baseline condition, was $r = 0.055$. To conclude, the computational contextualization model simulates meaningful datasets based on the individually optimized parameters. Although correlations between empirically observed data and their simulated counterpart are still rather modest, they may increase with improved cognitive context schemata, spatial bias distributions, and parameter optimization algorithms in future versions of the model. In any case, the current version of the model is already able to capture a certain proportion of the variance in behavioral patterns.