Aspects of Rational Argumentation

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M.Sc. Bruno Richter

aus Suhl

Tübingen

2022

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

Tag der mündlichen Qualifikation:	20.09.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Mandy Hütter
2. Berichterstatter:	Prof. Dr. Klaus Fiedler

For my family

Acknowledgments

This dissertation thesis summarizes research on a topic that is dear to me—rational argumentation. Throughout the years, I received great institutional and personal support that allowed me to finish my dissertation.

I thank the Studienstiftung des deutschen Volkes for granting me a scholarship that enabled me to focus on my studies. I still feel the privilege and responsibility that comes along with being an alumnus of this foundation. I am also grateful for the support grant from the Deutsche Gesellschaft für Psychologie as well as the graduation grant from the Gesellschaft für Kognitionswissenschaft, which helped me to finish my dissertation in its final stage. I am thankful to the University of Giessen, the Sapienza University of Rome, and the University of Tübingen, for providing stimulating intellectual environments, all of which contributed to making my ideas come true.

I also thank the colleagues and friends that I met along the way, from all of which I learnt something new. Importantly, I wish to express my gratitude to three intellectual mentors who had great positive impact on my work: First, I thank Prof. Dr. Markus Knauff. He warmly welcomed me in his lab, sparked my interest in rationality research, and instilled in me a humanist notion of education. Second, I thank Prof. Dr. Mandy Hütter for her consistent and reliable supervision, her good-hearted nature, and for always believing in me. Third, I thank Prof. Dr. Klaus Fiedler. He empowered me, boosted my confidence, and encouraged me to follow my own path.

I have dedicated this thesis to my family—this decision was easy.

Abstract

Rational argumentation is shaped both by logical norms and pragmatic principles. Normative accounts of rationality intend to understand rational argumentation through the lens of formal logic and propositional calculus. Descriptive accounts of rationality intend to understand rational argumentation through the lens of psychological mechanisms and the construction of meaning through linguistic rules. The current thesis argues against this dichotomous approach of logic versus pragmatics, and proposes some aspects for an integrated normative-descriptive model of rational argumentation. Based on a conceptual integration of relevance theory and the argumentative theory of reasoning, and operationalized by a conditional reasoning paradigm that incorporates counterarguments, I test the hypothesis that logical and pragmatic factors jointly predict the inferred conclusions during rational argumentation. Specifically, the conclusion endorsement during rational argumentation depends on an interaction between inference type of the conditional and linguistic mode of the counterargument. Throughout a series of three experiments, mixed model analyses, and meta-analyses, I find confirmatory evidence for my hypothesis. The interaction effect of inference type and linguistic mode is replicable, reproducible, and robust. I further identify relevance as an essential boundary condition of the effect, provide tentative evidence for its invariance across languages, and obtain response time data in order to better understand the underlying cognitive mechanisms. Taken together, the findings are highly indicative of an interplay of logical factors and pragmatic factors during rational argumentation. This suggests that the establishment of an integrated normative-descriptive model is the best way forward in order to achieve a comprehensive understanding of rational argumentation.

Keywords: rationality, argumentation, reasoning, logic, pragmatics, relevance

Zusammenfassung

Rationale Argumentation wird durch logische Normen und pragmatische Prinzipien geprägt. Normative Rationalitätsansätze beabsichtigen, rationale Argumentation durch die Linse formaler Logik und propositionalen Kalküls zu verstehen. Deskriptive Rationalitätsansätze beabsichtigen, rationale Argumentation durch die Linse psychologischer Mechanismen und die Konstruktion von Bedeutung durch linguistische Regeln zu verstehen. Die vorliegende Thesis argumentiert gegen eine dichotome Herangehensweise von Logik versus Pragmatik und schlägt einige Aspekte für ein integriertes normativ-deskriptives Modell rationaler Argumentation vor. Basierend auf einer konzeptuellen Integration der Relevanztheorie und der argumentativen Theorie des Denkens, sowie operationalisiert durch ein Paradigma des konditionalen Schlussfolgerns, welches Gegenargumente einbezieht, teste ich die Hypothese, dass logische und pragmatische Faktoren zusammenwirkend die gefolgerten Konklusionen während rationaler Argumentation vorhersagen. Im Besonderen hängt die Befürwortung der Konklusion während rationaler Argumentation von dem Inferenztyp des Konditionals und dem linguistischen Modus des Gegenarguments ab. Im Verlauf einer Serie von drei Experimenten, gemischter Modellanalysen und Metaanalysen finde ich konfirmatorische Evidenz für meine Hypothese. Der Interaktionseffekt von Inferenztyp und linguistischem Modus ist replizierbar, reproduzierbar und robust. Ich identifiziere überdies Relevanz als eine essenzielle Grenzbedingung des Effekts, liefere vorläufige Evidenz für dessen Invarianz über Sprachen hinweg und beziehe Antwortzeitdaten, um die zugrunde liegenden kognitiven Mechanismen besser zu verstehen. Insgesamt sind die Befunde hochgradig indikativ für ein Zusammenspiel von logischen Faktoren und pragmatischen Faktoren bei rationaler Argumentation. Dies legt nahe, dass die

Etablierung eines integrierten normativ-deskriptiven Modells der beste Weg nach vorne ist, um ein umfassendes Verständnis rationaler Argumentation zu erreichen.

Schlüsselwörter: Rationalität, Argumentation, Denken, Logik, Pragmatik, Relevanz

Table of Contents

1 Introduction

Rational argumentation is ubiquitous in human life. We use it to negotiate our interests in a complex world. The production and evaluation of rational arguments enables functional communication across a wide array of domains, such as law, politics, science, business, and our personal lives. It is therefore not surprising that rational argumentation constitutes a strong predictor for professional and personal success. This renders the study of rational argumentation a fascinating research topic that promises to lead to insights that will be of extraordinarily high value for individuals and society. Another inherent feature of rational argumentation that makes it particularly attractive as a research topic lies in the fact that rational argumentation is a cultural tool that exerts power without using power. Indeed, it is the most effective tool that homo sapiens developed for influencing others without using sheer force. It is precisely this quality of rational argumentation that Habermas (1981) referred to as *der zwanglose Zwang des besseren Arguments*. This German idiom expresses the idea that, as long as arguments and counterarguments can be expressed freely, the better argument will eventually prevail. However, what qualifies as a good argument? How does the interplay of arguments and counterarguments facilitate a form of argumentative reasoning that can be defined as rational? In line with the reflections that the early Wittgenstein (1922) expressed in the *Tractatus logico-philosophicus*, normative accounts of rational argumentation focus on the formal rules of logical reasoning and propositional calculus. In contrast, corresponding to the ideas that the late Wittgenstein (1953) articulated in the posthumously published *Philosophical Investigations*, descriptive accounts of rational argumentation study its psychological mechanisms and pragmatic principles. I believe that an erudite approach to rational argumentation requires the dialectical integration of both normative accounts and descriptive accounts for a more comprehensive understanding of the topic. My hope is

that the present thesis provides some *prolegomena* for a unified theory of rational argumentation.

2 Theoretical Background

Chapter 2 builds the theoretical foundation of the thesis. I will commence this chapter with a paragraph that intends to show the historical roots of the study of rational argumentation—pragmatics. Then, the first and second subchapter ought to describe the key theories informing this thesis, relevance theory and the argumentative theory of reasoning. The third subchapter provides an overview of conditional reasoning as a paradigmatic field and research methodology to study rational argumentation. The fourth subchapter outlines the influence of counterarguments in conditional reasoning. This will eventually lead to the derivation of the research rationale that underlies this thesis in the fifth and last subchapter of the theory section.

The historical roots of the scientific study of rational argumentation can be traced back to the emergence of pragmatics as a new field in the cognitive sciences, especially within linguistics. Pragmatics studies how contextual factors shape linguistic meaning and modulate the interpretation of arguments (Austin, 1962; Sperber & Wilson, 2002; Wearing, 2015; Wilson & Sperber, 2012a). Theoretical precursors and early works related to contemporary pragmatics research have their origin in philosophy. Morris (1938) described pragmatics as the scientific investigation of linguistic signs, interpreters of linguistic signs, and how signs and interpreters relate. In 1967, the philosopher Paul Grice offered groundbreaking new ideas in his William James lectures at Harvard University. This initiated a tremendous amount of new theorizing and research in the field. Grice (1975) introduced a new conceptual tool, which he coined conversational implicature. He distinguished between explicatures, information that was explicitly uttered, and implicatures, messages that were implicitly meant. Implicatures refer to the non-conventional meaning of an argument, which is

often determined by contextual factors of the speech act. He argued that rational argumentation is a fundamentally intentional activity: It enables communication and understanding by virtue of expressing intentions on the side of the sender as well as recognizing those intentions on the side of the recipient of an argument (Grice, 1957, 1969, 1982, 1989). Consequently, the expression of intentions by a sender and the processing of those intentions by a recipient facilitate rational communication. Both these parties are assumed to adhere to a cooperative principle and four maxims of communication, which are defined as follows by Grice (1975, pp. 45–46):

Cooperative Principle

Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.

Maxim of Quantity

1. Make your contribution as informative as is required (for the current purposes of the exchange).
2. Do not make your contribution more informative than is required.

Maxim of Quality

Supermaxim: Try to make your contribution one that is true.

1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence.

Maxim of Relation

Be relevant.

Maxim of Manner

Supermaxim: Be perspicuous.

1. Avoid obscurity of expression.
2. Avoid ambiguity.
3. Be brief (avoid unnecessary prolixity).
4. Be orderly.

Following the cooperative principle and the maxims of quantity, quality, relation, and manner promotes conversational rationality during argumentation, and their violation obstructs rational argumentation (Grice, 1989). His theory was and remains to be extremely successful and has inspired the development of new pragmatic theories. Highly influential Neo-Gricean theories of pragmatics stem from Gazdar (1979), Levinson (1983, 2000), Horn (1984, 1989, 2000, 2006), and Atlas (2005). However, a weak spot in Grice's theory is the third maxim. Whereas the other maxims are specified by supermaxims and/or submaxims, the maxim of relation simply states *be relevant* but other than that remains unspecified. It is precisely through this weak spot that a new theory was born—relevance theory.

2.1 Relevance Theory

Starting at Grice's (1975) legacy, Sperber and Wilson (1986, 1995) devised an innovative new theory that put the concept of relevance center stage (Gorayska & Lindsay, 1993). Relevance theory is an integrated theory of human cognition and communication (Wilson & Sperber, 1994). Note that relevance theory does not attempt to characterize the concept of relevance in linguistic terms (e.g., by means of a semantic analysis). This would neither be an expedient endeavor nor a particularly rewarding exercise because the word "relevance" is a vague notion, which can have

different meanings to different people. Moreover, it does not have a direct one-to-one translation in every human language. Instead, Sperber and Wilson (1986) argued:

> We do believe, though, that scientific psychology needs a concept which is close enough to the ordinary language notion of relevance; in other words, we believe that there is an important psychological property—a property of mental processes—which the ordinary notion of relevance roughly approximates, and which it is therefore appropriate to call relevance too, using the term now in a technical sense. What we are trying to do is to describe this property: that is, to define *relevance* as a useful theoretical concept. (p. 119)

According to this psychological definition, relevance conceptualizes a feature of inputs to cognitive processing. Such inputs can be perceived external stimuli or mentally manipulated internal representations (Blakemore, 2001; Carston, 1999; Wilson & Sperber, 1994, 2012a). An input is considered relevant to the degree that it combines with contextual information in order to achieve a useful cognitive output. Humans have a biologically rooted tendency for relevance maximization. This tendency is an evolutionary outcome of selection pressure toward the need to augment cognitive efficiency by concentrating on relevant input and ignoring or only superficially processing irrelevant input (Sperber & Wilson, 2002). This aptitude of the architecture of the human mind was a necessary condition for the genesis of higher-order cognitive competences and metarepresentational cognitive operations like the general ability to monitor and understand one's and other people's thoughts and intentions (Wilson & Sperber, 2002; Yus, 1998, 2010). The central tenet of relevance theory (Sperber & Wilson, 1986, 1995) is that relevance is governed by an interaction of an input's positive cognitive effects as well as its necessary processing effort. The greater the

attained positive cognitive effects, and the smaller the required mental effort, the greater is the input's relevance for an individual at a given point in time. Consequently, relevance maximization constitutes a function of facilitating positive cognitive effects whilst keeping processing costs reasonably low (Gibbs Jr. & Tendahl, 2006). This function is driven by two generic principles, which are specified as follows by Sperber and Wilson (1995, p. 260):

Cognitive Principle of Relevance

Human cognition tends to be geared to the maximization of relevance.

Communicative Principle of Relevance

Every act of ostensive communication communicates a presumption of its own optimal relevance.

The cognitive principle of relevance emphasizes the predisposition of the human mind to maximize relevance for attaining cognitive efficiency (i.e., positive cognitive effects and low processing effort). The communicative principle of relevance highlights an inherent property of communicative actions, namely that they transfer the assumption to be optimally relevant for the recipient. Two conditions must be met for this presumption to be successfully applied (Sperber & Wilson, 1995, p. 270):

Presumption of Optimal Relevance

1. The ostensive stimulus is relevant enough for it to be worth the addressee's effort to process it.
2. The ostensive stimulus is the most relevant one compatible with the communicator's abilities and preferences.

Based on the cognitive principle of relevance, the communicative principle of relevance, and the presumption of optimal relevance, it is claimed that the processing of arguments follows a relevance-guided comprehension heuristic (Sperber et al., 1995, p. 51):

Relevance-guided Comprehension Heuristic

1. Considering possible cognitive effects in their order of accessibility (i.e., following a path of least effort).
2. Stopping when the expected level of relevance is achieved (or appears unachievable).

This heuristic aids the recipient of the argument to reconstruct the meaning of the sender's argument. This task requires the recipient to place the argument into an appropriate context, utilize background knowledge, check source trustworthiness, monitor sender expertise, and employ expectations of relevance. The subtasks involved in the comprehension process are defined as follows by Wilson and Sperber (2012a, p. 13):

Subtasks of the Comprehension Process

1. Constructing an appropriate hypothesis about explicatures by developing the linguistically encoded logical form.
2. Constructing an appropriate hypothesis about the intended contextual assumptions (i.e., the implicated premises).
3. Constructing an appropriate hypothesis about the intended contextual implications (i.e., the implicated conclusions).

Wilson and Sperber (2012a) emphasize that these subtasks need not operate in a sequentially ordered fashion. Rather, they should be thought of as an online process that simultaneously generates hypotheses about explicatures, assumptions, and implications, based upon prior knowledge and other contextual cues, which in turn contribute to the revision or elaboration of the argument during the relevance-theoretic comprehension heuristic. The above specification of the relevance-theoretic comprehension process of rational argumentation indicates another property of relevance: Relevance is an intrinsically context-dependent property of arguments (or other cognitive inputs). An argument counts as relevant for a specific context if it exerts contextually related effects within it. (Sperber & Wilson, 1986). Put differently, relevance is not an absolute property of an argument, but rather a matter of degree relative to its context (Levinson, 1989; Wearing, 2015). Albeit relevance theory capitalizes on the cognitive layers of communication and argumentation, its social references are undeniable. A major assumption of relevance theory regards ostensive-inferential communication. This form of communication encapsulates two kinds of intentions—informative intentions and communicative intentions. Informative intentions represent the purpose of informing an audience about something. Communicative intentions represent the purpose of informing an audience of one's informative intention. Respective actions associated with those intentions of ostensive-inferential communication serve as a prime source of information about one's social environment during rational argumentation, which is why the social relevance of relevance theory is incontestable (Allott, 2013; Jary, 1998; Wilson & Sperber, 2002).

A major strength of relevance theory is that the relevance-theoretic framework combines theoretical clarity with sophisticated experimentation and as such constitutes an experimentally testable cognitive theory of reasoning, communication, and rational argumentation (Wilson & Sperber, 2002). A notable number of predictions on cognitive

and pragmatic performance have been deduced from relevance theory and were subjected to the empirical test. For instance, it has been shown that the relevance-theoretic comprehension procedure and the presumption of optimal relevance afford a causally plausible explanation for results obtained with the Wason selection task (Girotto et al., 2001; Sperber et al., 1995; Sperber & Girotto, 2003). Sperber et al. (1995) experimentally varied the cognitive effects available in carrying out the task as well as the necessary processing effort. The authors demonstrated that both factors influenced the selection of cards, which suggests that these factors are crucial for the comprehension process. Consequently, as evidenced by four experiments, the authors proposed a general and predictive explanation of the Wason selection task based on relevance theory, and found empirical support for their claims. Sperber et al.'s (1995) paper is an exemplary article showing the potential of relevance theory to produce specific, unequivocal hypotheses, how to test these hypotheses experimentally using a classic conditional reasoning paradigm, and how the obtained findings can foster the induction of novel implications associated with the theory itself. Furthermore, relevance theory has proven to be successfully applicable to and highly influential in other research areas (Yuan et al., 2019; Yus, 2010), such as theory of mind (e.g., Happé, 1993), humor (e.g., Yus, 2003, 2016), media discourse (e.g., Tanaka, 1994), discourse analysis (e.g., Ifantidou, 2014; Pilkington, 2000; Schourup, 2011; Tendahl & Gibbs Jr., 2008), politeness (e.g., Christie, 2007; Jary, 1998; Mazzarella, 2015; Ruytenbeek, 2019), translation studies (e.g., Díaz-Pérez, 2014; Gutt, 2000), and emotion research (e.g., Wharton et al., 2021).

At this point, I have introduced relevance theory (Sperber & Wilson, 1986, 1995) as one major theoretical building block of this thesis. I have shown that the psychological construct of relevance is a crucially important theoretical concept to understand both human cognition and communication at large. It is now necessary to

connect the general theory of relevance with a more domain-specific theory that directly addresses the specific scope of research of the present thesis, namely rational argumentation. Together, the first theoretical pillar—relevance theory—and the second theoretical pillar—the argumentative theory of reasoning—lay a solid foundation to build upon the research rationale of this thesis.

2.2 Argumentative Theory of Reasoning

In a stellar article entitled “Why do humans reason? Arguments for an argumentative theory”, published 2011 in the journal *Behavioral and Brain Sciences*, Hugo Mercier and Dan Sperber introduced a novel and highly innovative theory of rational argumentation—the argumentative theory of reasoning. The core thesis of their theory is the assertion that human reasoning evolved to produce and evaluate arguments (Mercier & Sperber, 2011). While others have already suggested earlier that the predominant function of reasoning may be argumentative in nature (Billig, 1996; Gonsrth & Perelman, 1949; Perelman & Olbrechts-Tyteca, 1969; Toulmin, 1958), these accounts were mostly based on introspective inquiry and approached the issue from a purely philosophical perspective. In contrast, Mercier and Sperber (2011) offer a fully spelled out theory that argues from a naturalistic and evolutionary perspective (see also Dessalles, 2007), and is supported by a plethora of empirical findings. Not only has the theory produced a myriad of new original research on the topic of rational argumentation, it is also able to account for past findings that were hard to make sense of before (Sperber & Mercier, 2012). The ingenious coup of the argumentative theory of reasoning is that it provides a subversive and compelling alternative to the mainstream notion that the main function of reasoning was to enhance individual cognition (e.g., Kahneman, 2003). Mercier and Sperber (2011) impressively demonstrate that humans actually perform quite poorly on tasks that represent this function; instead, human reasoning functions much better in

argumentative contexts. The influence of the idea that reasoning is for arguing became so influential in psychological science that its importance for the field can hardly be overstated.

The main assumption of the argumentative theory of reasoning is that human reasoning has evolved because it makes communication more effective and advantageous by means of rational argumentation, which, in turn, is fueled by two central mental operations: the production of arguments and the evaluation of arguments. The human capacity for argument production is designed in a way that fosters the generation of arguments that support the view of the arguer. This leads people to being prone to bias during the production of arguments because they mostly search for arguments that support their own standpoint. People tend to not be exactly exigent toward checking the validity of their own arguments. Hence, argument production underlies weak quality control. During the evaluation of others' arguments, however, people accept even challenging arguments, given these arguments are convincing. They do so because argument evaluation underlies a strict quality control—people only consider the arguments of others if these arguments are relevant for the issue at stake. Table 1 lists the characteristics of rational argumentation for the production of arguments and the evaluation of arguments, with respect to proneness to bias and quality control.

Given the inherent properties of rational argumentation as it manifests itself in argument production and argument evaluation, the argumentative theory of reasoning derives distinct predictions for these two characteristics. On the one hand, when reasoners produce arguments, they are biased and lazy. This is due to the fact that reasoners typically try to convince the other side in an argumentative context. Consequently, they search for, generate, and communicate arguments that support their own opinion. During argument production, the application of counterarguments

Table 1. Characteristics of rational argumentation.

	Argument Production	Argument Evaluation
Bias	Biased: people mostly produce reasons for their side	Unbiased: people accept even challenging reasons, if they are strong enough
Quality Control	Lazy: people are not very exigent toward their own reasons	Demanding: people are convinced only by good enough reasons

Note. Adapted from Mercier and Sperber (2017, p. 235).

would be counter-productive; it would serve the interlocutor and weaken the arguments that support the standpoint the arguer intends to propose or defend, respectively. Thus, argument production is characterized by a confirmation bias, also known as myside bias. It is defined as "[…] the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" (Nickerson, 1998, p. 175). While the confirmation bias has generally been considered a flaw of human reasoning, Mercier and Sperber (2011) assert that whether the effects of the confirmation bias are adverse or beneficial crucially depends on the context. During solitary thinking and within groups whose members hold largely converging beliefs, confirmation bias is responsible for poor performance, which is often connected to motivated reasoning (Kunda, 1990), groupthink (Janis, 1972), and group polarization (Myers & Lamm, 1975). By contrast, in argumentative contexts in which divergent views are exchanged, where arguments and counterarguments flow freely, and when the interlocutors share the common desire to find the truth (or at least approximate it), then the confirmation bias is constructive because it facilitates an efficient division of cognitive labor under such conditions. Every member of the debate will produce the strongest arguments

that represent their own view. Hence, the confirmation bias is a feature rather than a flaw during argument production (Mercier & Landemore, 2012).

However, the argumentative theory derives an asymmetrically different prediction for the second characteristic of rational argumentation, namely argument evaluation. During the evaluation of arguments, reasoners should be unbiased and demanding with respect to monitoring the received arguments for quality. Reasoners tend to evaluate arguments objectively since they intend to assess the relevance of the argument for the issue at hand. If they evaluate a counterargument as weak, not convincing, or irrelevant, they will reject it. In contrast, if they evaluate a counterargument as strong, convincing, or relevant, they will integrate it in order to revise prior beliefs and adjust their decision-making. The degree to which a reasoner objectively engages in argument evaluation depends on two factors—the extent of dialog and the extent of conflict. Dialog connotes the amount of arguments; the more arguments available, the higher the chances for unbiased evaluation. Dialog can occur internally (e.g., via private deliberation) and externally (e.g., via public debate). Conflict connotes the ratio of the convergence and divergence of arguments; the more arguments and counterarguments oppose each other, the higher are chances for rational outcomes (Mercier, 2016a; Mercier & Sperber, 2017). The demanding but unbiased process of argument evaluation explains why rational argumentation enables the best argument to carry the day in argumentative contexts (e.g., Moshman & Geil, 1998; Trouche et al., 2014).

Taken together, the human capability to produce and evaluate arguments constitutes the main function of why humans reason in the first place. Mercier and Sperber (2017) claim that the traditional intellectualist view of reasoning, namely that reasoning serves to help individuals gain knowledge and make better decisions on their own, does not sufficiently explain all nuances of reasoning. Instead, they propose

a new perspective from an interactionist view of reasoning, arguing that reasoning mainly contributes to the pursuit of social interaction goals by justifying oneself and convincing others through good argumentation. Hence, the argumentative theory of reasoning is inherently social (Sperber, 2001). In fact, the emphasis on the social dimension of reasoning, the idea that reasoning is a profoundly social mental activity, can be traced back to ancient Greek philosophy, where Socrates applied his method of Elenchus (i.e., maieutics) to promote critical thinking and stimulate new insights in his students, and where the citizens of the polis gathered at the agora to debate current political questions and related issues of societal relevance, thus establishing the procedural foundations of liberal democracy and the constitutional preconditions of modern Western society. Historical key figures of psychology have also highlighted the close connection between rationality and sociality. Piaget (1928) stated that:

> The social need to share the thought of others and to communicate our own with success is at the root of our need for verification. Logical reasoning is an argument which we have with ourselves, and which reproduces internally the features of a real argument. (p. 204)

Similarly, the sociocultural tradition of Vygotsky (1978, p. 57), who stated that “every function in the child’s cultural development appears twice: first, on the social level, and later, on the individual level; first *between* people (interpsychological), and then *inside* the child (intrapsychological)”, aligns with the idea that rational thought is always embedded within a social or argumentative context. According to Vygotsky (1962), reasoning is deliberative in function and dialogic in structure. In other words, reasoning is an internalized analogue of interpersonal discourse (Mercier & Sperber, 2011). The everyday social practice of argumentation within family and other cultural

entities (e.g., school, work, peer group, neighborhood, etc.) is the ontogenetic trigger point that instantiates the formation of internalized dialogic structures that foster the development of individual reasoning skills that serve an ultimately deliberative-argumentative function.

So far, it has been made clear that the central function of reasoning is to engage in rational argumentation. However, one might wonder why and how this function came into existence and was encoded into the human mind at all. Why did homo sapiens develop the extraordinarily complex skill to produce arguments and evaluate the arguments of others? What was the adaptive value of doing so? The argumentative theory of reasoning provides an evolutionary rationale to tackle these questions: Reasoning evolved because it increases human fitness. Reasoning empowers a speaker to argue for his claim, and a hearer to evaluate these arguments. Thus, reasoning improves the quantity and the epistemic quality of shared information (Darmstadter, 2013). Shared information, in turn, is essentially important for survival since humans heavily rely on communication (Dawkins & Krebs, 1978; Krebs & Dawkins, 1984). Humans had the need to monitor the arguments they encountered in order to detect potential cheaters during cooperative acts. Scanning the communicated arguments helped to protect humans against manipulation attempts, misinformation, lies, and unjust exchange (Dessalles, 2011; Mercier, 2013a; Sperber, 2001). Producing weak arguments for your case as well as evaluating others' arguments without the necessary effort was much more costly in the archaic times of our ancestral past. The immediate lethal consequences of weak argumentative reasoning shaped the human mind into a cognitive device whose mental operations are specifically tuned to optimally solve the problems related to such scenarios. Therefore, the competence to reason and argue rationally became an adaptive tool that proved remarkably successful in boosting fitness. Due to selection pressure, rational argumentation

prevailed over other features that turned out to be less advantageous for survival. Mercier (2011a) concedes that evolutionary rationales can be speculative. However, an evolutionary rationale can be used to deduce predictions that can only follow from this specific rationale. These predictions can then be tested empirically against predictions that stem from competing hypotheses. When the empirical results correspond to the predictions of the evolutionary rationale, and when at the same time no evidence can be found that confirms the predictions of the competing hypotheses, then it becomes likely that the mechanism or phenomenon under investigation has an evolutionary foundation (Mercier, 2011a). Indeed, the evolutionary rationale of the argumentative theory of reasoning prompts predictions that are consistent with data from many fields (e.g., Mercier, 2013a, 2013b).

The natural predisposition of humans for rational argumentation is tightly geared to the concept of epistemic vigilance (Sperber et al., 2010). The tremendous value of shared communication for human fitness is incontestable. Information can be so beneficial to humans, and misinformation can be so detrimental, that humans must have evolved a mechanism that keeps them constantly vigilant towards the epistemic value of the arguments they produce and evaluate. Epistemic vigilance consists of two functions—trust calibration and coherence checking. Trust calibration is responsible for checking personal properties of the interlocutor, for example source reliability, trustworthiness, reputation, credibility, expertise, and social status. Coherence checking refers to monitoring the content the interlocutor utters, for example argument strength, validity, plausibility, utility, and correspondence with prior beliefs. Together, both functions—trust calibration and coherence checking—enable humans to benefit from rational argumentation. Therefore, epistemic vigilance is an indispensable presupposition for rational argumentation (Mercier & Sperber, 2017; Sperber et al., 2010).

Further support for the evolutionary rationale of the argumentative theory of reasoning comes from cultural psychology. Traditionally, cultural psychology examines the cultural practices within cultures and investigates similarities and differences between various cultures (Henrich et al., 2010; Nisbett, 2003; Nisbett et al., 2001; Norenzayan et al., 2002; Peng & Nisbett, 1999). It has often been argued that the members of individualistic cultures, which are mostly found in Western societies, are predominantly adapted to an analytic thinking style that enhances debate and discursive argumentation. In stark contrast, the members of collectivistic cultures, which are commonly located in Eastern societies and also among traditional indigenous tribes, mainly adopted a holistic thinking style that fosters social harmony and dialectical argumentation. However, this is not to say that people from collectivistic cultures or tribal communities lack the ability of producing arguments for their side and objectively evaluating arguments from others. Of course, modern Western societies might excel in the practice of rational argumentation because they have been training this particular cultural practice for thousands of years. In hindsight, ancient Greece is often identified as the starting point of this development. But in fact, the tradition of rational argumentation—i.e., objective evaluation of arguments, fair consideration of counterarguments, and approaching a question from different perspectives—is deeply ingrained in Jewish philosophy (Putnam, 2008) and thus even older than often assumed. Judaism is one of the oldest cultures that uses scripts to pass on knowledge, defines education as being of utmost importance in their value system, and cultivated the practice of debating over religious texts (e.g., Talmudic commentary) for thousands of years. As Jews moved to Europe, they established Sephardic communities in Southern Europe and Ashkenazic communities in Eastern Europe. Hence, Jewish philosophy has greatly contributed to the high priority of rational argumentation in Western societies. Albeit Asian societies of the Far East and indigenous communities

might not have emphasized the role of rational argumentation in their cultural practices as much, the evolutionary rationale of the argumentative theory of reasoning implies the hypothesis that selection pressure has shaped the ability to produce and evaluate arguments into the human mind. Therefore, a feasible prediction is that the argumentative function of reasoning should be found universally across all cultures (as I argued above, maybe in some to a higher or lesser extent than others, respectively, due to specific cultural practices). Indeed, the empirical evidence suggests that rational argumentation, as it ought to function according to the argumentative theory of reasoning, is a cross-culturally robust and universal skill of the human mind (Mercier, 2011a). For example, Mercier et al. (2016) tested the hypothesis that Easterners might not share the benefits of rational argumentation in a Japanese sample. In Experiment 1, participants had to solve the Wason selection task, a standard logical problem, individually and in groups. They performed significantly better on the task when working in groups. In Experiment 2, participants were asked to give numerical estimates for the weights of various animals, first individually, then after learning of another participant's estimate, then after discussing the estimates with the other participant, and finally individually again. The benefits of the discussion with the other participant became visible in the participants' final individual estimates. Results from both experiments suggest that the overall positive efficacy of rational argumentation, evoked by discussing with another participant, is comparable to that observed in studies conducted with Western samples. Similarly, Castelain et al. (2016) tested whether the members from a traditional indigenous Maya population from Guatemala reason according to the predictions of the argumentative theory of reasoning. Participants had to perform a series of volume conservation tasks (Dasen, 1972; Piaget & Inhelder, 1974), first individually (pre-test), then in groups (test), and finally individually again (post-test). Consistently across two studies, the performance

significantly increased during the test phase, and remained robust in the post-test. These results show that, apart from WEIRD samples (i.e., Western, educated, industrialized, rich, and democratic samples), members from preliterate indigenous communities reason according to the argumentative theory of reasoning: reasoning improves with discussion; argument production is marked by confirmation bias; argument evaluation is effective; people distinguish between strong and weak arguments; and people are only convinced by strong arguments.

More empirical evidence is supportive of the argumentative theory of reasoning as it shows that rational argumentation improves reasoning performance on a variety of tasks, including logical, mathematical, and inductive problems (Laughlin, 2011; Moshman & Geil, 1998; Trouche et al., 2014), moral reasoning (Mercier, 2011b), scientific reasoning (Johnson, 2011; Mercier & Heintz, 2014), work-related tasks (Mercier, 2011c), and school tasks (Mercier, 2016b; Mercier et al., 2017; Slavin, 1995; Smith et al., 2009). The developmental evidence corroborates these findings, too (Mercier, 2011d). Children of three years are already able to engage in argumentation (Stein & Albro, 2001; Stein & Bernas, 1999; Stein & Miller, 1993). Preschoolers are already capable of detecting argumentative fallacies like circular reasoning (Baum et al., 2008). Neuroscientific investigations into the neural basis of argumentative reasoning demonstrate that argumentative reasoning is associated with enhanced activity in the medial prefrontal cortex (mPFC; Prado et al., 2020).

All things considered, the argumentative theory of reasoning (Mercier & Sperber, 2011) provides a highly comprehensive theoretical framework that is extremely successful in generating original research to test its underlying assumptions (Mercier, 2016a). The theory is particularly refreshing because it draws a positive picture of human rationality by re-interpreting reasoning phenomena that have otherwise been considered flaws and turning them into strengths (Mercier & Sperber,

2017). Moreover, the argumentative theory of reasoning has proven highly applicable in several fields of empirical conduct investigating reasoning. One major field of investigation in this regard concerns conditional reasoning. The next chapter is devoted to illuminating the working mechanisms and functions of arguments in conditional reasoning.

2.3 Conditional Reasoning

Conditional reasoning is a form of deductive reasoning. Unlike other forms of reasoning, for instance inductive reasoning and abductive reasoning, deduction is traditionally defined as an inference process during which a valid conclusion *necessarily* follows from premises that are assumed to be true (Manktelow, 2012). While deduction also includes syllogistic reasoning and relational reasoning, it is the specific characteristics of conditional reasoning that suggest its particular suitability to study rational argument. Conditionals, the propositional entities of conditional reasoning, are ubiquitous in rational argumentation. They occur in all human languages (Comrie, 1986) and are the most extensively studied propositions in the psychology of reasoning (Byrne & Johnson-Laird, 2009; Evans & Over, 2004; Oaksford & Chater, 2010). Many decades of conceptual work and empirical research have produced a highly diverse cluster of theories on conditional reasoning. These theories differ in their main assumptions, normative standards, and descriptive findings (Knauff & Gazzo Castañeda, 2021; Knauff & Spohn, 2021). Yet, they can be roughly divided into two camps: classical logic approaches and probabilistic approaches. Classical logic approaches include mental proofs (Inhelder & Piaget, 1958; Rips, 1994), natural logic (Braine & O'Brien, 1998), pragmatic schemata (Cheng & Holyoak, 1985), mental model theory (Johnson-Laird, 1983, 2006; Johnson-Laird & Byrne, 1991, 2002), dual-process theory (Evans, 2003, 2008, 2019; Evans & Stanovich, 2013; Klauer et al., 2010; Singmann et al., 2016; Stanovich, 2011; Verschueren et al., 2005),

argumentative theory (Mercier & Sperber, 2017), evolutionary adaptation (Cosmides, 1989), atmosphere (Woodworth & Sells, 1935), and matching bias (Evans, 1972). Probabilistic approaches include Bayesian rationality (Hahn, 2014; Oaksford & Chater, 2007), suppositional theory (Edgington, 1995; Evans, 2007; Evans & Over, 2004; Evans et al., 2005), probability logic (Pfeifer, 2013; Pfeifer & Kleiter, 2010), computational rationality (Gershman et al., 2015; Tenenbaum et al., 2006), and ranking theory (Skovgaard-Olsen, 2016a; Spohn, 2012).

Conditional inference consists of a major premise, a minor premise, and a conclusion. The major premise is a conditional statement of the form ‘if p, then q’, where p denotes the conditional’s antecedent and q denotes the conditional’s consequent. The minor premise as well as the validity of the conclusion are contingent on the inference type. There are four inference types, which are: modus ponens (MP), denial of the antecedent (DA), affirmation of the consequent (AC), and modus tollens (MT). They are expressed as follows:

$$\text{MP} = \frac{p \rightarrow q,\ p}{\therefore q}, \tag{1}$$

$$\text{DA} = \frac{p \rightarrow q,\ \neg p}{\therefore \neg q}, \tag{2}$$

$$\text{AC} = \frac{p \rightarrow q,\ q}{\therefore p}, \tag{3}$$

$$\text{MT} = \frac{p \rightarrow q,\ \neg q}{\therefore \neg p}, \tag{4}$$

where Equation 1 formalizes modus ponens, Equation 2 formalizes denial of the antecedent, Equation 3 formalizes affirmation of the consequent, and Equation 4

formalizes modus tollens. Modus ponens and modus tollens are normatively correct inferences that are logically valid; the conclusion is necessarily true given the premises are true. Denial of the antecedent and affirmation of the consequent are normatively incorrect inferences that are logical fallacies; the conclusion is not necessarily true even if the premises are true.

The cognitive psychologist Peter Wason designed a clever paradigm to study conditional reasoning—the Wason selection task (Wason, 1966). It is presumably the most investigated experimental paradigm in the field. The impact it exerted on the psychology of reasoning as a whole can hardly be overstated. Stenning and van Lambalgen (2008) called it “the mother of all reasoning tasks” (p. 44). Figure 1 shows the standard Wason selection task.

The basic idea of the task is that the rule represents a major premise of the form ‘if p, then q’. Each card represents a minor premise: ‘E’ represents p, ‘K’ represents ¬p, ‘2’ represents q, and ‘7’ represents ¬q. Hence, the major premise together with each one of the minor premises represent all four inference types. The rule plus ‘E’ represents modus ponens. The rule plus ‘K’ represents denial of the antecedent. The rule plus ‘2’ represents affirmation of the consequent. The rule plus ‘7’ represents modus tollens. By implication, participants should turn over the card with the vowel ‘E’ to test whether there is an even number on the other side, which would be the valid modus ponens inference. Also, they should turn over the card with the *uneven* number ‘7’ to test whether there is *not* a vowel on the other side, which would be the valid modus tollens inference. The other cards, ‘K’ and ‘2’, should be ignored because turning them over would test the invalid denial of the antecedent and affirmation of the consequent inferences.

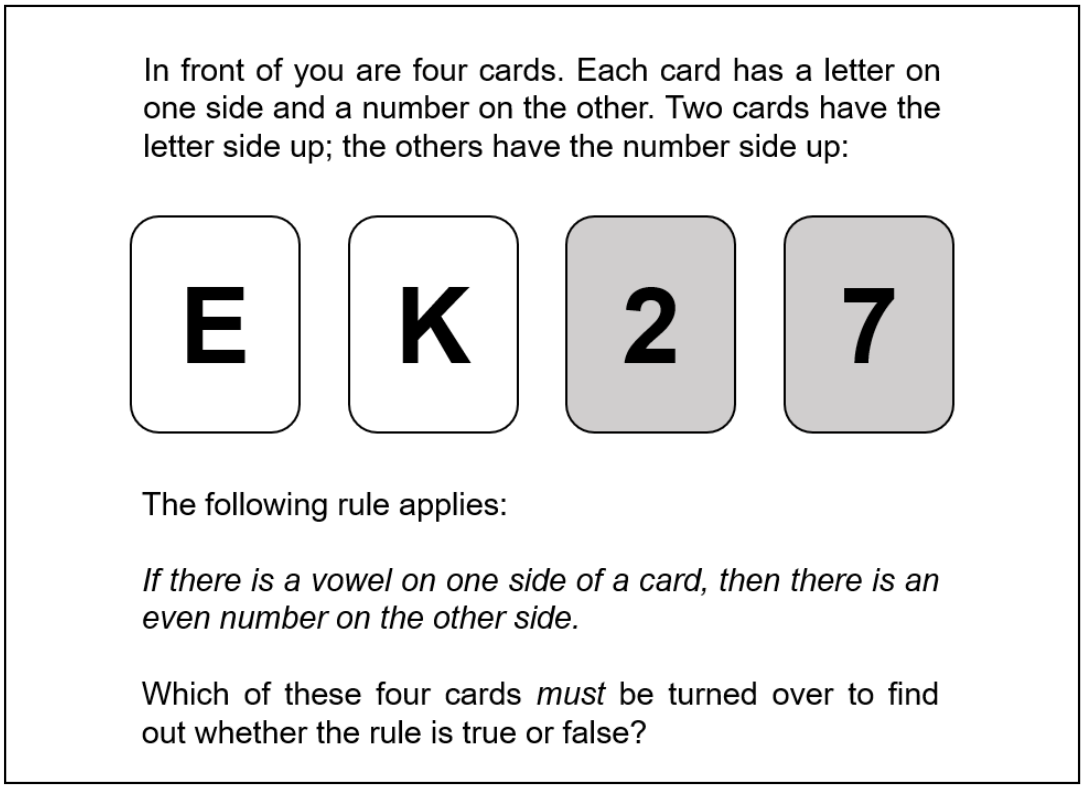

Figure 1. The standard Wason selection task (adapted from Mercier & Sperber, 2017, p. 212).

The first original research on the Wason selection task was conducted by its inventor himself. Wason (1968) found that only one of 34 participants correctly selected the cards representing p and ¬q. Most participants either picked the p card alone, or they picked the p card and the q card. Wason and Johnson-Laird (1972) replicated this pattern, and it has not changed since (e.g., Evans et al., 1993; Klauer et al., 2007). In a detailed meta-analysis reviewing 34 studies with 845 participants in total and conducted over a time span of over 25 years, Oaksford and Chater (1994) report relative frequencies of 89% for choosing the p card, 16% for choosing the ¬p card, 62% for choosing the q card, and 25% for choosing the ¬q card. This extremely robust pattern demonstrates that people perform very good at drawing the modus ponens inference, whereas the modus tollens inference seems to be more difficult and is therefore endorsed less frequently. A possible explanation for this phenomenon might

be the disposition of humans to select evidence during hypothesis testing in a way that seeks to corroborate the hypothesis, rather than searching for potential counterexamples that would refute the hypothesis (Ragni et al., 2018). Epistemically, this is not a wise strategy—no amount of confirming instances can prove a hypothesis ultimately true; however, according to critical rationalism, a single falsifying instance can prove it wrong!

The selection task has launched much research on rational argumentation with conditionals throughout the years. It stimulated research with variations of the standard form of the task as well as the introduction of new experimental paradigms to study conditional reasoning. Obviously, the present treatise cannot treat all of it. However, Schroyens et al. (2001) provide a meta-analytic review that clearly shows the general response patterns of conditional reasoning. The authors synthesized results from conditional reasoning experiments including over 700 participants. The presented conditional arguments contained abstract materials to minimize the influence of general knowledge. Figure 2 shows the meta-analytically combined relative frequencies for the endorsement of each of the four inference types. As can be seen, people do not simply respond according to the norms of formal logic. If they did, they would always endorse modus ponens and modus tollens, but never endorse denial of the antecedent and affirmation of the consequent. The different endorsement ratings of modus ponens and modus tollens are particularly puzzling. Why is modus ponens more frequently endorsed than modus tollens? Given that both inferences are logically valid, should they not be equally endorsed? Any psychological theory of conditional reasoning must therefore explain why modus tollens is less frequently accepted than modus ponens (Manktelow, 2012).

Notwithstanding that some aspects of conditional reasoning have yet to be solved, it is clear that pertinent tests of conditional reasoning have already helped us

to better comprehend the boundary conditions and mechanistic processes of rational argumentation. However, a conundrum that remains refers to the meaning of counterarguments for rational argumentation. The next chapter highlights the role of counterarguments for rational thought and argumentation, reviews the literature on conditional reasoning with counterarguments, and reports the state of the art of its empirical investigation.

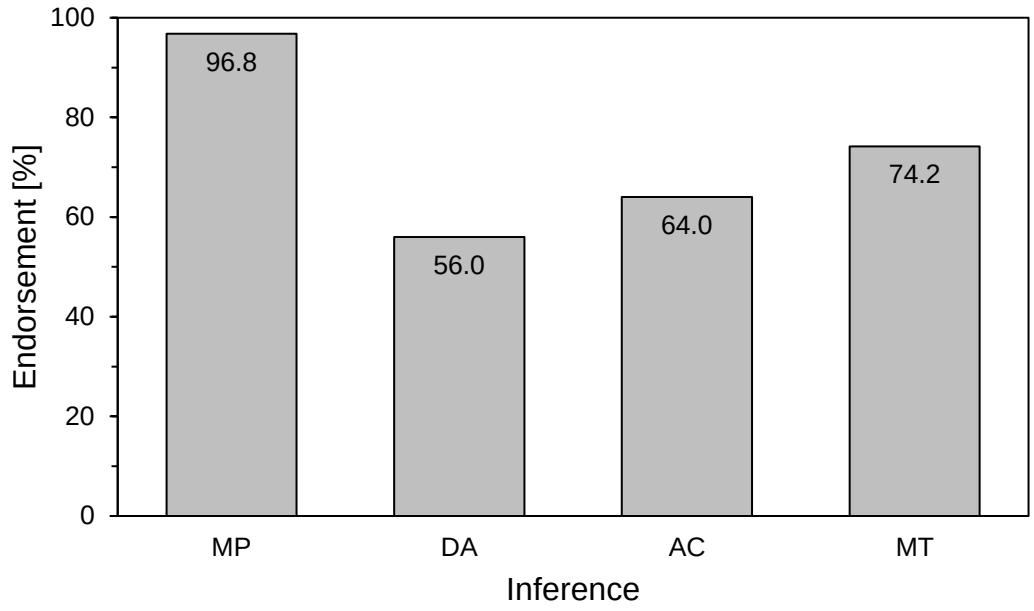

Figure 2. Results adapted from Schroyens et al. (2001, pp. 168–172). MP = modus ponens; DA = denial of the antecedent; AC = affirmation of the consequent; MT = modus tollens.

2.4 Conditional Reasoning with Counterarguments

Counterarguments are paramount for rational thought in general. They can help us to avoid an omnipresent bias—the confirmation bias. This bias describes the tendency of humans to adopt a positive test strategy of confirmation rather than a negative test strategy of disconfirmation, whether it concerns hypothesis testing in scientific investigation or belief consolidation in everyday reasoning (e.g., Kappes et

al., 2020; Klayman & Ha, 1987; Lord et al., 1979; Nickerson, 1998; Rajsic et al., 2015; Wason, 1960). This test strategy is not optimal because it induces systematic errors and epistemic inefficiencies in the long run. Interestingly, the confirmation bias (or myside bias) shows very little to no relation to intelligence and other cognitive ability measures (Stanovich & West, 2007, 2008; Stanovich et al., 2013). This, again, corroborates that everyone can succumb this cognitive bias. However, there is an effective strategy that helps to reduce the proneness to confirmation bias and mitigates its deleterious consequences—confronting yourself or being confronted with counterarguments. Indeed, as has been stated in the chapter explicating the argumentative theory of reasoning (Mercier, 2016a; Mercier & Sperber, 2011, 2017), humans are less prone to confirmation bias during argument *evaluation* (as opposed to argument production). Dibbets and Meesters (2020) showed that counter-attitudinal information alters pre-existing beliefs and accompanying patterns of information search in the direction of being less prone to confirmation bias. Drummond and Fischhoff (2019) demonstrated that the priming of individuals' scientific reasoning skills reduces confirmation bias when directly instructed to apply those skills to the task at hand. Hernandez and Preston (2013) showed both quasi-experimentally (i.e., using naturally occurring groups) for political ideology attitudes and experimentally (i.e., using randomly assigned and systematically manipulated groups and conditions) for positivity toward a court defendant that the experienced difficulty (i.e., disfluency) in processing arguments reduced the confirmation bias by promoting careful, analytic processing. Notably, this effect did not occur when participants were under cognitive load, suggesting that free cognitive resources must be available to overcome the confirmation bias. Shehab and Nussbaum (2015) also provide evidence indicating that the processing of counterarguments and its effects on rational argumentation depend on cognitive load. Using the argument-counterargument integration paradigm

(Nussbaum, 2008), participants either (a) constructed claims that minimize the disadvantages of an alternative or (b) weighted refutations that weaken an argument by arguing that there are more important values at stake. Weighing complex refutations was associated with more cognitive load than constructing claims, arguably because disparate elements in working memory must be coordinated during the former. Taken together, the here reported empirical findings strongly suggest that counterarguments exert a remarkable influence on rational argumentation. Specifically, the evaluation of counterarguments leads to more complex and refined cognitive processing, which in conclusion promotes an objective, balanced, bilateral, and impartial integration of arguments and counterarguments.

A germane method to study the impact of counterarguments on rational argumentation is conditional reasoning. Specifically, the key question motivating the present research is: How do people engage in rational argumentation in the light of counterarguments during conditional reasoning? My core assumption states that humans reject otherwise valid conclusions in the light of new evidence inferred from an evaluation of relevant counterarguments. In conditional reasoning with counterarguments, people still integrate the initial premises in their inference process. However, counterarguments are thought to indicate specific circumstances that imply the inference that some conclusions are not coercive anymore (cf. Gazzo Castañeda & Knauff, 2021a; Gazzo Castañeda et al., 2016). In other words, conditional reasoning with counterarguments is a non-monotonic inference process. Rational agents retract the consequences of previous arguments when convincing counterarguments enter the inferential space (Elio & Pelletier, 1997). In the psychology of reasoning literature, conditional reasoning with counterarguments is frequently referred to as defeasible reasoning. This nomenclature is inspired by deliberations on this type of inference process in philosophy (e.g., Spohn, 2020) and research in artificial intelligence (e.g.,

Pollock, 1987, 2001). Traditionally, two disparate types of counterarguments (defeaters) are distinguished in defeasible reasoning: alternatives and disablers. Alternatives are alternative reasons for the consequent that are not the antecedent. They signify that the antecedent is *sufficient but not necessary* for the consequent to occur. Hence, alternatives are crucial for affirmation of the consequent and denial of the antecedent inferences. By contrast, disablers are additional preconditions that prevent the antecedent from leading to the consequent. They indicate that the antecedent is *necessary but not sufficient* for the consequent. Disablers are relevant for modus ponens and modus tollens inferences.

The first empirical study on conditional reasoning with counterarguments dates back to the 1980s. Rumain et al. (1983) found that alternatives reduce the endorsement of affirmation of the consequent inferences as well as denial of the antecedent inferences. This effect was named the suppression effect. Byrne (1989) demonstrated that the suppression effect also applies to disablers. The presentation of different sorts of disablers (exceptions, counterexamples, statements expressing conditions of impossibility) reduced the endorsement of modus ponens and modus tollens inferences. The magnitude of the suppression effect rises as the number of counterarguments increases (Cummins, 1995; Cummins et al., 1991), as the strength of counterarguments increases (Chan & Chua, 1994), and as the counterarguments become more certain (Stevenson & Over, 1995). To date, numerous studies have been conducted demonstrating that counterarguments influence rational argumentation during conditional reasoning (e.g., Cummins et al., 1991; Demeure et al., 2009; De Neys et al., 2003a, 2003b; Markovits & Potvin, 2001). Typically, the experimental sequence of one trial begins with the presentation of the major premise, followed by the minor premise. Then, in the counterargument condition, a counterargument subsequently follows the minor premise. Finally, the conclusion is presented; it is either

shown as a statement or phrased as a question. The examples below illustrate the paradigmatic task, where p denotes the antecedent, q denotes the consequent, and c denotes the counterargument:

Example 1

If John studies hard, then he passes the exam. | if p then q

He passes the exam. | q

John cheats on the exam. | c

John studies hard. | p

Example 2

If John studies hard, then he passes the exam. | if p then q

John studies hard. | p

John suffers from insomnia. | c

He passes the exam. | q

Example 1 represents an affirmation of the consequent inference. Typically, the introduction of the counterargument c implies that the antecedent p is a *sufficient but not necessary* precondition for the consequent q to occur. John might pass the exam not because he studies hard, but because he cheats on the exam. Therefore, participants endorse the conclusion that John studies hard less frequently. This is reflected by decreased acceptance of the affirmation of the consequent inference type. This very logic is executed for the denial of the antecedent inference type, too. Example 2 represents a modus ponens inference. Typically, the introduction of the counterargument c implies that the antecedent p is a *necessary but not sufficient*

precondition for the consequent q to occur. Even though John studies hard he might not pass the exam, because he suffers from insomnia (and is therefore tired at the exam). Therefore, participants endorse the conclusion that he passes the exam less frequently. This is reflected by decreased acceptance of the modus ponens inference. The same logic applies to the modus tollens inference.

Two different paradigms exist to study the effects of conditional reasoning with counterarguments as exemplified above: an overt paradigm and a covert paradigm. The overt paradigm was first introduced by Rumain et al. (1983). In the overt paradigm, counterarguments are explicitly presented. When disablers are presented, participants accept fewer modus ponens conclusions and fewer modus tollens conclusions. Likewise, when alternatives are shown, participants accept fewer affirmation of the consequent conclusions and fewer denial of the antecedent conclusions (e.g., Byrne, 1989; Byrne et al., 1999; Stevenson & Over, 1995). The covert paradigm was first introduced by Cummins et al. (1991). In the covert paradigm, counterarguments are not explicitly presented but implicitly present in the premises. The experimental design is as follows: First, one group of participants are shown a set of conditionals. Their task is to generate as many counterarguments (disablers and alternatives) as possible. Depending on how many counterarguments were generated, the conditionals are then subdivided into those having (1) many disablers and many alternatives, (2) many disablers and few alternatives, (3) few disablers and many alternatives, and (4) few disablers and few alternatives. Finally, these conditionals are embedded into the four inference types (modus ponens, modus tollens, affirmation of the consequent, denial of the antecedent) and presented to a second group of participants. The task of this new group of participants, which never saw the actual counterarguments, is to indicate how strongly or not they endorse the respective conclusions. The general finding is that conclusion endorsement varies as a function of the number of presented

counterarguments in the overt paradigm or as a function of the number of generated counterarguments in the covert paradigm, respectively. Participants endorse modus ponens and modus tollens conclusions to a lesser degree when conditionals have many instead of few disablers. Likewise, they accept affirmation of the consequent conclusions and denial of the antecedent conclusions to a lesser degree when conditionals have many instead of few alternatives (e.g., Cummins, 1995; De Neys et al., 2003a, 2003b; Gazzo Castañeda & Knauff, 2018).

Notably, conditional reasoning with counterarguments heavily depends on content and background knowledge (e.g., see Beller & Spada, 2003; De Neys et al., 2003a, 2003b; Dieussaert et al., 2005; Evans & Over, 2004; Johnson-Laird & Byrne, 2002; Oaksford & Chater, 2007). Experimentally, content effects have been detected by contrasting the endorsement of conclusions for inference tasks with familiar versus unfamiliar content (Cummins, 1995; Gazzo Casteñeda & Knauff, 2021b; Markovits, 1986), and by having lay people versus experts compare the validity of domain-specific counterarguments (Gazzo Casteñeda & Knauff, 2016a). The studies on content effects in conditional reasoning with counterarguments yield a consistent pattern of results: A lack of domain-specific knowledge leads participants to accept more conclusions, regardless of the inference type. Instead, possessing background knowledge renders participants aware of the specific situations whose occurrence would circumvent the conclusion. Thus, counterarguments affect the perceived *sufficiency* and *necessity* relationships between the antecedent part and the consequent part of a conditional argument. Counterarguments in the form of disablers make the antecedent part less sufficient for the consequent part; consequently, fewer modus ponens conclusions and modus tollens conclusions are accepted. Counterarguments in the form of alternatives make the antecedent part less necessary for the consequent part; consequently, fewer

affirmation of the consequent conclusions and denial of the antecedent conclusions are accepted (Cummins, 1995; Thompson, 1994, 1995).

Other important factors that affect conditional reasoning with counterarguments are associative strength, number, frequency of occurrence, working memory, and context. Associative strength refers to the link between the conditional premises and the counterargument in declarative memory. Some counterarguments are more strongly associated with prior knowledge of how to prevent an otherwise valid conclusion than others (Quinn & Markovits, 1998). These counterarguments have a higher associative strength in declarative memory, are thus more easily retrieved, and eventually more likely to defeat an otherwise valid conclusion (De Neys et al., 2003a; Vadeboncoeur & Markovits, 1999). Additionally, the higher the number of available counterarguments, the more probable it is that a conclusion will be defeated (De Neys et al., 2003b). Interestingly, Geiger and Oberauer (2007) showed that it is not only the associative strength and the number of counterarguments that count. Sometimes, there can be many potential counterarguments but their frequency of occurrence is low (and vice versa). Geiger and Oberauer (2007) experimentally disentangled the effects of number versus frequency of occurrence and found that frequency of occurrence predicts conclusion endorsement more accurately than the mere number of potential counterarguments. Regarding working memory, Toms et al. (1993) already provided first evidence for the notion that the limited capacity of working memory interferes with the cognitive demands associated with conditional reasoning tasks. De Neys et al. (2005) specifically examined how working memory contributes to the retrieval (or inhibition) of counterarguments during conditional reasoning. In a first experiment, the authors showed that people with high working memory spans rejected logically invalid inferences more frequently than people with low working memory spans, whereas people with low working memory spans rejected logically valid inferences more

frequently than people with high working memory spans. In a second experiment, a secondary task demanding executive attention was imposed. This increased cognitive load by consuming additional working memory capacity. The results indicate that working memory resources are used for the retrieval of stored counterarguments from memory. Lastly, another factor that must not be neglected can have an essential impact on conditional reasoning with counterarguments, namely context. Vadeboncoeur and Markovits (1999) demonstrated that the activation and retrieval of counterarguments can be inhibited when participants are explicitly instructed to focus on the logical necessity of the conditionals. Even the response format of the dependent measure can influence conditional reasoning with counterarguments. When using a scaled response format, the number of counterarguments affects conclusion endorsement. However, when using a dichotomous response format, a single counterargument is already enough to reject a conclusion (Markovits et al., 2010).

A specific set of context effects that is of particular interest to me refers to pragmatic modulations of conditional reasoning with counterarguments. When humans monitor pragmatic context, they infer intended meanings by assuming that speech acts convey only relevant information, which is one of the most astonishing features of human communication and rationality (Frank & Goodman, 2012). However, the preconditions, operating principles, and implied consequences of pragmatics have often been neglected in the study of conditional reasoning with counterarguments, leaving an enigmatic research gap in the relevant literature with many remaining questions yet to be answered. The central goal of the present thesis is to illuminate some aspects of rational argumentation by virtue of studying the pragmatic principles informing, underlying, and predicting conditional reasoning with counterarguments. I hope this will help to capture some of the richness of rational argumentation in pragmatic context. To this end, the next chapter identifies the open research gaps in

the relevant literature, formulates the research questions I aim to answer, and summarizes the research objective of the present thesis.

2.5 Research Rationale

The rationale motivating the present investigation, as outlined before, is based on a theory integration of relevance theory (Sperber & Wilson, 1986, 1995) and the argumentative theory of reasoning (Mercier & Sperber, 2011, 2017). Relevance theory provides a viable account of how humans engage in conditional reasoning. Sperber et al. (1995) demonstrated that relevance theory provides a general and predictive explanation of how people reason when they select evidence for testing a conditional rule in the Wason selection task (Wason, 1966). Participants infer testable consequences from the conditional rule; they stop testing as soon as the resulting interpretation of the rule meets their expectation of relevance. Importantly, participants' expectation of relevance varies with inference type and context of the conditional rule, and, consequently, so does their performance. These results confirm that relevance constitutes a key concept in the study of conditional reasoning in particular and rational argumentation in general. It also shows that relevance theory offers experimentally testable predictions and thus provides an empirically oriented theory of rational argumentation (Wearing, 2015; Wilson & Sperber, 2002). Furthermore, and most importantly, the findings from Sperber et al. (1995) show that rational argumentation transcends the limits of a purely linguistic analysis of the formal grammar (syntax) and explicit contents (semantics) in conditional reasoning. Instead, a viable account of rational argumentation requires pragmatics. This is due to the fact that natural language and reasoning does not operate like a computer language, where logical syntax and necessary semantic input are sufficient preconditions for a functional computation. Instead, conditional reasoning processes, which energize rational argumentation, largely depend on pragmatic inference. It is pragmatic inference that

allows humans to utilize any *relevant* premise that the context provides, assess its believability, and draw a conclusion from it with a certain degree of conviction. Hence, the consideration of the pragmatic level advances the quest to comprehend the operating principles of rational argumentation in conditional reasoning (Evans & Over, 2004). However, while relevance theory explains how humans draw conclusions based on the presumption of optimal relevance in conditional reasoning, it is still unclear how the theory's assumptions map onto conditional reasoning with counterarguments. Specifically, when testing conditional rules, it is unclear how reasoners differentially respond to counterarguments that satisfy the presumption of optimal relevance versus counterarguments that violate the assumption of optimal relevance. It is precisely this research gap that must and can be filled by a theory integration of relevance theory and the argumentative theory of reasoning. The application of relevance theory in conjunction with the argumentative theory of reasoning to conditional reasoning with counterarguments generates testable and falsifiable predictions of how rational argumentation in pragmatic contexts functions. While, on the one hand, relevance theory pinpoints the concept of relevance as decisive aspect of rational argumentation, the argumentative theory of reasoning, on the other hand, provides the missing piece to predict how exactly relevance modulates conditional reasoning with counterarguments as a function of pragmatic context variables. According to the argumentative theory of reasoning, people reason better and more sophisticated within an argumentative context (Sperber & Mercier, 2012). For example, people tend to make correct use of modus tollens arguments when eager to attack alternative views (Pennington & Hastie, 1993), whereas they oftentimes fail to do so in standard reasoning tasks that lack an argumentative context (Evans et al., 1993). Notably, this argumentative context is not solely established externally by means of public debate. Rather, argumentative context may also be given internally via thoughtful deliberation

and careful analysis and weighting of all arguments and counterarguments given (Mercier & Landemore, 2012). The crucial aspect of argumentative context, whether it is established externally or internally, is that it must create some sort of conflict and (intraindividual or interindividual) dialog to be constructive (Mercier, 2016a). Accordingly, rational argumentation does also take place within the *individual* mind by involving an anticipatory and imaginative communicative framing of arguments and counterarguments (Sperber et al., 2010). Hence, even the solitary thinker can engage in rational argumentation, provided that he is exposed to arguments that generate conflict. In the case of conditional reasoning, this condition is met when counterarguments are presented. Here, the magnitude of conflict is contingent on the quality of a counterargument, which is, in turn, essentially determined by its relevance. A relevant counterargument leads people to envisage not the neglect of the conditional itself, but the denial of the consequent given its antecedent (Byrne & Johnson-Laird, 2009; Johnson-Laird et al., 2009).

An integration of relevance theory and the argumentative theory of reasoning provides a meta-theoretical framework to identify research gaps in the study of rational argumentation. This paragraph deals with the research questions and research objective to help close those gaps. The research questions address the idea that rational argumentation depends on pragmatic context. I ask four questions:

Research Questions

1. What is the role of inference type for rational argumentation?
2. How does the linguistic mode of a counterargument affect rational argumentation?
3. How do inference type and linguistic mode jointly predict rational argumentation?
4. When and how does relevance modulate the joint influence of inference type and linguistic mode in rational argumentation?

Considering the research questions formulated above, I herewith summarize the research objective of my doctoral thesis:

Research Objective

The research objective of the present thesis is to study the pragmatic modulation of rational argumentation in conditional reasoning with counterarguments.

Surprisingly, previous research on the pragmatic modulation of rational argument is relatively scarce. Roberge (1978) had adult participants perform a propositional reasoning task. He experimentally manipulated the semantic content in which the logical rule was embedded, the linguistic form of the logical rule, and the polarity of the major premise of the logical arguments. He found that all three factors influenced reasoning performance, thereby yielding first suggestive evidence for the impact of pragmatic variables in reasoning. Evans (1983) showed that the widely documented matching bias in conditional reasoning depends on linguistic factors. Specifically, he found that the introduction of explicit negatives in a conditional rule significantly reduces the matching bias compared with the normal usage of affirmative instances of the rule. Hilton et al. (1990) demonstrated that conditional reasoning can be modulated by the presence of contextual assertions that affirm or deny the existence of alternative causes in the causal field of a conditional. Other authors have equally argued that rational argumentation is predominantly pragmatic in nature (for a review, see Evans, 2002). It is this psychological phenomenon of context variables and pragmatic nuances influencing rational argument that is referred to as *pragmatic modulation* (Johnson-Laird & Byrne, 2002; Levinson, 1983). While the findings reported above should be regarded *prima facie* evidence for the important role of pragmatics in rational argumentation, Stenning and van Lambalgen (2004) rightly

predicted that the interest in pragmatics will revive and stimulate a new wave of research into conditional reasoning. Indeed, it has only been research in very recent years that suggests a revival of interest in pragmatics within the fields of conditional reasoning, rationality, and argumentation research (e.g., see Peloquin et al., 2020). Gazzo Castañeda and Knauff (2016b) examined the role of the modal auxiliaries ‘should’ and ‘will’ in deontic reasoning with conditionals. For modus ponens inferences that were phrased with the modal ‘should’, people selected conclusions based on their own sense of justice, whereas this effect was attenuated when the deontic conditional was phrased with the modal ‘will’. However, modus tollens inferences remained unaffected by modal auxiliaries. In another study, Gazzo Castañeda and Knauff (2019) demonstrated across three experiments that the specificity of terms influences conditional reasoning, too. Inferences comprising specific terms were endorsed more frequently as were inferences entailing unspecific terms. Gazzo Castañeda and Knauff (2021b, 2021c) also demonstrated that this effect remains robust even without prior knowledge about an inference, for example when participants are faced with counterintuitive conditionals (i.e., rules describing the opposite of what is to be expected from everyday experience) and with arbitrary conditionals (i.e., rules without an obvious causal link between the antecedent and the consequent). Both the early and the current studies reported in the paragraph above indicate that pragmatic factors critically influence processes of rational argumentation in conditional reasoning. However, the state of the art also makes it clear that quite a few research gaps remain to be bridged, especially with respect to the role of counterarguments in rational argumentation—and their modulation by inference type, linguistic mode, and relevance.

Firstly, referring to the first research question of the present thesis, I set out to clarify the role of inference type for rational argumentation. Here, I will focus on the two

valid inference types modus ponens and modus tollens. These inference types are especially suited to investigate rational argumentation because their logical validity sets a normative standard for the measurement of rational argumentative performance during conditional reasoning tasks. Starting from the seminal work of Byrne (1989) using the suppression task, by now it has been widely documented that people endorse the conclusion of the modus ponens inference with a considerable degree of certainty. In contrast, people seem to have more difficulties with accepting the conclusion of the modus tollens inference. Several explanations were postulated by various schools of thought to account for this imbalance of acceptance rates between modus ponens and modus tollens. Schroyens et al. (2001) argue that effects of inferential negation are responsible for this imbalance. Modus tollens entails negations, in its explicit and manifest form of the minor premise (i.e., ¬q), in its implicit and latent inference during cognitive processing, and in its explicit and manifest conclusion (i.e., ¬p). These inferential negations put additional cognitive load on the reasoner, which creates interference with the conditional inference process and therefore distorts the inference itself. Consequently, people are sometimes led to falsely conclude p instead of ¬p. An alternative explanation for the imbalance between modus ponens and modus tollens relies on the directionality of human reasoning. Stenning and van Lambalgen (2005) pose that the causal direction of the modus ponens inference is predictive in nature. People are better adapted to such types of easy forward reasoning, which is why reasoning performance of modus ponens inferences is high. In contrast, the direction of the modus tollens inference is diagnostic by nature. This kind of complex backward processing is less common in human inference, which consequently accommodates for the lower performance in modus tollens reasoning tasks. However, it should be noted that the alternative explanation of Stenning and van Lambalgen (2005) should not be regarded as a competitor of, but rather as a concomitant to the explanation of

inferential negation postulated by Schroyens et al. (2001). Based on these findings and its algorithmic explanations, I hypothesize that inference type plays an important role for rational argumentation. Specifically, I predict that modus ponens yields higher endorsement than modus tollens in rational argumentation as operationalized by conditional reasoning with counterarguments.

Secondly, referring to the second research question of the present thesis, a further research goal is to clarify how counterarguments affect rational argumentation. In particular, I am interested in how exactly the variation of the linguistic mode (i.e., the phrasing in subjunctive mode versus indicative mode) of the counterargument affects endorsement. The basic idea is that counterarguments in subjunctive versus indicative mode carry different pragmatic connotations that can manipulate the believability of the conditional and the counterargument itself, respectively. In line with this reasoning, Gazzo Castañeda and Knauff (2021a) state that the consideration of counterarguments can be moderated by the introduction of certain words that question the believability of the conditional or the counterargument. George (1997) presented conditional reasoning tasks that were either phrased in the traditional form ‘if p then q’ or with an additional ‘very probable’ or ‘not very probable’ before the conditional. He found that this certainty manipulation has influenced participants’ acceptance of the conclusion. Stevenson and Over (1995) elicited similar effects by introducing doubt in the counterargument. One example conditional was ‘If John goes fishing, he will have a fish supper’. The respective counterargument was ‘If John catches a fish, he will have a fish supper’. However, when doubt in the counterargument was introduced by saying that ‘John is always lucky’, then participants rejected fewer conclusions. These findings show two things: Counterarguments reduce endorsement of otherwise valid conditional inferences, and this reduction can be modulated by subtle pragmatic variations in the conditional or the counterargument. Corresponding to these insights,

I hypothesize that counterarguments affect rational argumentation. Specifically, I predict that counterarguments reduce conclusion endorsement, both when presented in subjunctive mode and in indicative mode.

Thirdly, referring to the third research question of the present thesis, another research aim is to examine how precisely inference type of the conditional rule and linguistic mode of the counterargument jointly predict rational argumentation. I consider this a serious research gap in the literature. To the best of my knowledge, nobody has addressed this issue before. This lack of knowledge calls for being filled by a systematic empirical investigation. So, how might inference type and linguistic mode interact? Consistent with Johnson-Laird and Ragni (2019), I hypothesize that reasoners represent counterarguments as mental models of possibilities. Each counterargument possesses a possibility tag, whose salience is modulated by the linguistic mode of the counterargument. Since the indicative counterargument expresses a more certain possibility, its respective possibility tag is more salient. By contrast, the subjunctive counterargument expresses a less certain possibility, which is why its respective possibility tag is less salient. This, consequently, decisively modulates the degree to which reasoners endorse modus ponens versus modus tollens conclusions. In general, people have the tendency to reduce conflict by reasoning from inconsistency to consistency (Gawronski & Strack, 2012; Johnson-Laird et al., 2004). The introduction of counterarguments, of course, creates conflict, which reasoners try to dissolve as they differentially integrate the counterarguments' possibility tags as a function of inference type. For the case of modus ponens, the less salient possibility tag of the subjunctive counterargument leads reasoners to question its certainty. Instead, the more salient possibility tag of the indicative counterargument increases the certainty of its validity. Following this line of reasoning, I predict that subjunctive counterarguments reduce conclusion endorsement to a lesser extent than

indicative counterarguments for modus ponens inferences. However, I predict that this pattern should reverse for modus tollens inferences. This prediction of a reversed response pattern for modus tollens inferences, namely that indicative counterarguments reduce conclusion endorsement to a lesser extent than subjunctive counterarguments, is supported by the mismatch principle (Johnson-Laird et al., 2004). Elio and Pelletier (1997) provided suggestive evidence for the assumption that it is more probable to reject the conclusion of a modus ponens inference than to reject the conclusion of a modus tollens inference given a high degree of conflict between the conditional and the counterargument—and this high degree of conflict is exactly what the indicative counterargument with its highly salient possibility tag elicits. An indicative counterargument that is in high conflict with the modus ponens inference has the form ¬q, and so it also conflicts with the consequent clause of the conditional rule. In contrast, an indicative counterargument that is in high conflict with the modus tollens inference has the form p, which does *not* conflict with the antecedent clause of the conditional rule (see also Revlin et al., 2001). In summary, I hypothesize that effects of inferential negation and strategies that individuals use to cope with inconsistencies (e.g., see Hasson & Johnson-Laird, 2003) predict an interaction of inference type and counterarguments' linguistic mode in rational argumentation during conditional reasoning with counterarguments. Specifically, I predict that modus ponens conclusions are endorsed more frequently than modus tollens conclusions when the counterargument is subjunctive. Conversely, I predict that modus ponens conclusions are endorsed less frequently than modus tollens conclusions when the counterargument is indicative.

Fourthly and lastly, referring to the fourth research question of the present thesis, I aim to show that relevance functions as an important boundary condition of rational argumentation. I suspect that the relevance of a counterargument is

predetermined by its semantic content put in relation to the pragmatic context of the entire reasoning task, including major premise, minor premise, and, needless to say, the counterargument itself. Sperber et al. (1995) argue that the pragmatic context determines in which direction relevance is being envisaged, and directs the reasoner's attention and inference in that direction. If the content of the counterargument is unmediatedly connected to the focal context of the reasoning problem, then the counterargument carries a high value of relevance. If this is not the case, its relevance value is low, indicating irrelevance. Arguably, the predisposition to monitor the environment (e.g., argumentative context) for relevant cues (e.g., good counterarguments) evolved because it is important for humans to differentiate important from unimportant input before inferential integration starts. Sorting out irrelevant information before its actual integration in the inference process helps alleviate the entire processing phase and thus saves constrained cognitive resources (Sperber & Wilson, 1996). Instead, if a counterargument is relevant, humans are apt to carefully muster and evaluate it (Cacioppo & Petty, 1979; Edwards & Smith, 1996). Therefore, I hypothesize that relevance affects the influence of counterarguments on rational argumentation. Specifically, I predict a three-way interaction between inference type, linguistic mode, and relevance. I presume that this three-way interaction is established by virtue of relevance functioning as a boundary condition of the two-way interaction pattern of inference type and linguistic mode. In detail, this means that I expect the interaction between inference type and linguistic mode to be established via counterarguments' relevance. If, however, this boundary condition of relevance is not fulfilled, then the interaction breaks down. Note that this prediction is backed up by research showing that the relevance of a counterargument is paramount for defining its epistemic strength (Darmstadter, 2013; Hornikx & Hahn, 2012; Ransom et al., 2016). Relevance evokes reflection (Sperber & Mercier, 2012), which leads to a

fine-grained and nuanced processing of counterarguments within their pragmatic contexts of inference type and linguistic mode. It is the designated objective of the next chapter to put my predictions to the empirical test.

3 Empirical Evidence

Chapter 3 constitutes the empirical and statistical part of this thesis. First, I report three experiments. Experiment 1 is an original study and the first empirical test to address the main research questions of this thesis. Experiment 2 is a direct replication of the previous experiment in another language, with a different participant pool, and among a new external setting. Experiment 3 is a second replication as well as an extension of the previous experiments since it implements a nested study design that includes a new factor expected to function as a crucial moderator of the observed effects. Second, I report an extensive mixed-model analysis of all findings of the three experiments in order to test whether the empirical findings remain robust when the trial-to-trial variability in observations within participants is statistically accounted for. Third, I report a meta-analysis over all experiments to obtain combined effect size estimates for the observed effects. The function of the replications, the mixed-model analysis, and the meta-analysis is to empirically and statistically corroborate the robustness of the findings.

All computer programs, power calculations, data, analysis scripts, and modeling code of this thesis are documented and stored in a private OSF project under the persistent URL https://osf.io/3dm2j. I provide access to all files upon reasonable request. For all three experiments, I report how I determined sample size, all data exclusions (if any), all manipulations, and all measures. Prior to the experiments, participants gave written informed consent to the procedures. All procedures were in adherence with the ethical standards of the Declaration of Helsinki (World Medical Association, 2013) regarding the treatment of human participants in research.

3.1 Experiment 1

Experiment 1 served as a first original test to study the interplay of inference type, counterarguments, and the phrasing of the presented counterarguments with respect to their linguistic mode. I was especially interested in how exactly this interplay affects the degree to which a specific conclusion is endorsed, that is how strongly or not do participants agree that a certain conclusion must be inferred based on the combination of the preceding premises.

First, the impact of the mere inference type on conclusion endorsement has been intensively studied (e.g., Evans, 1977; Kern et al., 1983; Marcus & Rips, 1979; Markovits, 1988; Taplin, 1971). Although both modus ponens inferences and modus tollens inferences are logically valid and should be equivalently endorsed with a truth value of 1 from the normative standpoint of propositional logic (Copi, 1982), it is a broadly documented and well-established finding that modus ponens inferences and modus tollens inferences generally yield differential endorsement ratings. Specifically, modus ponens inferences are endorsed more frequently and more strongly than modus tollens inferences (for reviews see Evans, 1982; Wason & Johnson-Laird, 1972). Albeit modus tollens is a valid form of inference, people seem to find it to be a more difficult one. They are generally far more willing to infer modus ponens as opposed to modus tollens (e.g., Oaksford et al., 2000).

Second, in the psychology of reasoning it has long been neglected that humans do not only rely on the rules of formal logic during reasoning (Politzer & Bourmaud, 2002), despite the fact that the importance of this notion has long been recognized by mathematicians (Adams, 1975; Adams & Levine, 1975; Rescher, 1976; Suppes, 1966) and philosophers (Pollock, 1987). However, a strong theory of conditional reasoning must integrate the notion that logically valid conclusions can be withdrawn in the light of new evidence—for instance, by virtue of the introduction of counterarguments. The

classic study conducted by Byrne (1989) was the first to demonstrate that valid inferences can be suppressed by the presentation of counterarguments. This effect, which she coined the suppression effect, has been replicated and validated numerous times (e.g., Bonnefon & Hilton, 2004; De Neys et al., 2003a, 2003b; Johnson-Laird & Byrne, 1994; Thompson, 1994). The suppression effect describes the psychological phenomenon that reasoners tend to reject otherwise valid conclusions when they are confronted with an additional premise (i.e., some form of additional information) that elicits uncertainty in regards to the sufficiency and necessity of the causal link between antecedent and consequent of the conditional (Cummins, 1995; Cummins et al., 1991). These additional premises, which I refer to as counterarguments, are also termed defeaters or disablers in the relevant literature.

Third, and most interesting, more recently there have been attempts to study the role of semantic and pragmatic factors in conditional reasoning (Bonnefon & Villejoubert, 2007; Gazzo Castañeda & Knauff, 2016b, 2019, 2021c; Manktelow & Fairley, 2000). Albeit these studies provided first insights into how linguistic aspects can modulate conditional and defeasible reasoning, they are limited to the extent that they focused on the manipulation of linguistic elements of the conditional itself and did not include experimental variations of linguistic modalities of the counterargument. The role of linguistic mode in the consideration of counterarguments in inference tasks has thus not received much attention yet. To fill this gap in the research literature, a major aim of this study was therefore to investigate the influence of linguistic mode of the counterarguments, that is whether counterarguments are formulated in subjunctive mode or indicative mode, on conclusion endorsement. Importantly, I was particularly eager to understand how inference type and linguistic mode of the counterarguments interact. I assumed different response patterns in endorsement ratings for inference tasks with subjunctive counterarguments versus indicative counterarguments

depending on whether these inferences were modus ponens or modus tollens inferences. My reasoning was as follows: When people have to infer a conclusion based on a conditional but at the same time are presented a counterargument that raises doubt whether or not the conclusion still holds, they will search in semantic memory for further counterarguments, that is further exceptions to the rule, in order to validate their final decision to reject an otherwise valid conclusion. Indeed, De Neys et al. (2002, 2003a, 2003b) showed that reasoners tend to engage in a semantic search process for counterexamples to find further evidence that supports the implication to not endorse a certain conclusion that would be formally valid based on classical logic but is defeated given the counterargument. I claim that for subjunctive counterarguments, reasoners do not (or at least not with the same effort) engage in a search process for further counterarguments because they question the counterargument in the first place as the phrasing in subjunctive mode induces uncertainty regarding its validity. Consequently, reasoners should give higher endorsement ratings for modus ponens inferences versus modus tollens inferences with counterarguments in subjunctive mode. However, I presume that the pattern of responses should reverse when the counterarguments are presented in indicative mode. I argue that indicative mode activates the search process for other counterarguments. This search process, however, will be inhibited for modus tollens inferences since they require higher cognitive load. Thus, the additional processing effort related to modus tollens burdens the search process for more counterarguments. Consequently, less counterarguments are retrieved for modus tollens compared with modus ponens when the presented counterargument is phrased in indicative mode. This should lead to lower endorsement ratings for modus ponens as opposed to modus tollens inferences because more counterarguments could be retrieved during the search process in semantic memory.

To the best of my knowledge, this is a novel prediction that has never been empirically tested before. To this end, in Experiment 1 participants were given conditional reasoning tasks in which I manipulated inference type (modus ponens vs. modus tollens) and the mode of counterarguments (none vs. subjunctive vs. indicative). Participants' task was to indicate their degree of endorsement for the respective conclusion of each task. In addition, I assessed response times as a secondary measure. Note that I used response latencies as exploratory measure and thus did not state specific a-priori hypotheses regarding this dependent variable.

3.1.1 Hypothesis

I tested whether inference and mode influence endorsement ratings. I expected that both factors have an impact on endorsement ratings. First, I predicted that overall, modus ponens inferences display higher endorsement ratings than modus tollens inferences. Second, I predicted that overall, counterarguments reduce endorsement ratings. Third, and most importantly, I hypothesized that the interaction of inference and mode predicts endorsement ratings. Specifically, while for subjunctive mode endorsement ratings are higher for modus ponens inferences as opposed to modus tollens inferences, this relationship reverses for indicative mode. For indicative mode, endorsement ratings are lower for modus ponens inferences as opposed to modus tollens inferences. Finally, I explored response times to investigate the degree to which this second measure substantiates the findings obtained via the assessment of the endorsement ratings.

3.1.2 Method

This section gives a detailed description of the sample characteristics, study design, materials, and procedure of Experiment 1.

Participants. I conducted an a-priori power analysis with G*Power (Faul et al., 2007, 2009). Given a type I error of $\alpha = .05$ and assuming a power of $1 - \beta = .80$, 19

participants are sufficient to reliably detect a medium effect size of Cohen's f = .25 (Cohen, 1988; Ellis, 2010). In total, I collected data from N = 20 participants. A corresponding sensitivity analysis revealed that 20 participants are sufficient to reliably observe a minimum detectable effect (MDE) of Cohen's f = .24, given a type I error of α = .05 and assuming a power of 1 - β = .80. The final sample consisted of German university students of different majors who were between 19 and 29 years old (M = 22.35, SD = 3.36; 6 male). The first language of all participants was German. No participant had prior expertise in formal logic. Participants were recruited by sending a circular email via the email server of the University of Giessen, Germany. They were paid or received course credit for their participation.

Design. The experiment followed a 2 (inference: modus ponens vs. modus tollens) × 3 (mode: none vs. subjunctive vs. indicative) within-participants design. Dependent measures were endorsement rating and response time.

Materials. I created 36 scenarios in text form serving as reasoning tasks for Experiment 1. The scenarios concerned unequivocal, ordinary events from everyday life. Every scenario consisted of four stimuli: major premise, minor premise, counterargument, and conclusion. The major premise was always an if-then-clause, consisting of the antecedent (p) and the consequent (q). The minor premise expressed p for the "modus ponens" condition and ¬q for the "modus tollens" condition. The counterargument was either absent for the "none" condition, formulated in subjunctive mode for the "subjunctive" condition, or formulated in indicative mode for the "indicative" condition. The conclusion was phrased as question that asked for the respective response in the "modus ponens" condition and the "modus tollens" condition. All instructions (Appendix A1), stimuli (Appendix A2), and response formats (Appendix A3) of Experiment 1 are appended.

Procedure. After participants were welcomed and signed the informed consent form, they were seated in front of a computer screen in an individual lab room. The experiment was implemented and conducted using the program OpenSesame (Mathôt et al., 2012). I was present during the instructions and the practice trials to clarify potential questions before the main experiment started. First, participants were instructed about their task. They were informed that they will be presented reasoning tasks and that each reasoning task consists of several sentences. The first sentences are written in black font and include an if-then-clause, a fact, and possibly a counterargument. The last sentence is a question written in red font that asks for a conclusion. Participants were instructed to answer this question on a 7-point Likert scale using the numbers 1 to 7 written on green stickers, which were placed in one row on the number bar of the keyboard. Participants had to press the respective key to give their response. Key 1 represented “no, in no case”, while key 7 represented “yes, in any case”. For the counterbalanced version of the experiment, the scaling of the response format was reversed. Accordingly, key 1 represented “yes, in any case”, whereas key 7 represented “no, in no case”. Participants were instructed to provide responses based on how they would respond to these scenarios in everyday situations. Participants were informed that the sentences will appear sequentially and that they can proceed from one sentence to the next one by pressing the space bar. They were further informed that they can take a short break between tasks if needed and that they can continue by also pressing space bar. Participants were asked to read every reasoning task carefully since the sentences may also involve negations. If no questions remained, participants were asked to press space bar to start the practice phase of the experiment. The practice phase comprised three practice trials in order to familiarize the participants with their task. The practice trials were presented fully randomized. After the practice phase ended, participants were asked whether they

have any remaining questions. As soon as all questions (if any) were clarified, I left the room and the participants started the main phase of the experiment by pressing space bar. For each of the six experimental conditions, six scenarios were presented, resulting in 36 trials in total. The 36 trials were presented sequentially in fully randomized order. Each trial was presented once. Trials were given self-paced and without time restrictions. Participants proceeded from one trial to the next one by pressing space bar. Each trial comprised five stimuli that were presented sequentially in fixed order: a fixation point, a major premise, a minor premise, a counterargument (if applicable), and a conclusion. The conclusion was presented until the participants gave their response using the 7-point Likert scale on the keyboard. Stimuli appeared self-paced and without time restrictions. Participants proceeded from one stimulus to the next one by pressing space bar. Each stimulus was shown at the screen center of the monitor. The screen background was white. Participants received no feedback after giving a response. One half of the participants conducted the experiment with the default scaling of the response format, and the other half of the participants conducted the experiment with the counterbalanced scaling of the response format. Participants were alternately assigned to the default version or the counterbalanced version, respectively. After the main phase of the experiment ended, participants were thanked and asked to inform the experimenter. I then documented participants' demographic data (i.e., sex, age, occupation, major, nationality, first language, and expertise in formal logic) using a paper-pencil questionnaire. Lastly, the participants were debriefed, compensated, thanked, and dismissed.

3.1.3 Results

All participants met the inclusion criteria (i.e., German as first language, no expertise in formal logic) and completed the experiment without the occurrence of technical errors. Thus, no data had to be excluded. Data were processed and analyzed

using the statistical software R (Version 4.2.0; R Core Team, 2022) and JASP (Version 0.16.1; JASP Team, 2022). Practice trials were not included in the analysis. I conducted three important data transformations that refer to the correct coding of the endorsement ratings: (1) responses were inverted for the counterbalanced response format, (2) responses were inverted for modus tollens inferences because low/high values represent high/low endorsement, and (3) responses were linearly transformed from a 1-7 scale to a 0-6 scale in order to facilitate the readers' comprehension of the data visualizations. Data were processed to fit a wide format. I aggregated both endorsement ratings and response times across single trials for each participant and each experimental condition. Accordingly, I computed individual mean endorsement ratings and individual mean response times for each of the six experimental conditions.

For both dependent measures, I conducted a 2 (inference: modus ponens vs. modus tollens) × 3 (mode: none vs. subjunctive vs. indicative) repeated-measures ANOVA as omnibus test. In case of a significant interaction in the omnibus test, I separately conducted a 2 (inference: modus ponens vs. modus tollens) × 2 (mode: none vs. subjunctive) repeated-measures ANOVA, a 2 (inference: modus ponens vs. modus tollens) × 2 (mode: none vs. indicative) repeated-measures ANOVA, and a 2 (inference: modus ponens vs. modus tollens) × 2 (mode: subjunctive vs. indicative) repeated-measures ANOVA. For all computed repeated-measures ANOVAs, if Mauchley's test indicated that the assumption of sphericity is violated, then adjusted degrees of freedom and *p*-values based on the Greenhouse-Geisser correction (Greenhouse & Geisser, 1959) are reported. In case of significant interaction effects for the 2 × 2 repeated-measures ANOVAs, I calculated paired samples *t*-tests as post-hoc comparisons in order to further analyze the exact nature of the differences. I applied Holm's method (Holm, 1979) to adjust *p*-values for multiple post-hoc comparisons. I report Bayes factors for all analyses. For the repeated-measures

ANOVAs, I report BF_{incl}, which assesses the change from prior inclusion odds to posterior inclusion odds. It can be interpreted as the quantitative amount of evidence in the data for inserting a predictor into the model term. For the paired samples *t*-tests, I report BF_{10}, which indicates the relative evidence of H_1 over H_0 given the data. In addition, for the post-hoc comparisons of all significant 2 × 2 interaction effects, I provide a Bayesian sequential analysis representing how the evidence for H_1 or H_0, respectively, changes with increasing sample size.

Endorsement Ratings. The omnibus test, a 2 (inference: modus ponens vs. modus tollens) × 3 (mode: none vs. subjunctive vs. indicative) repeated-measures ANOVA, revealed a significant main effect for inference, $F(1, 19) = 5.49$, $p = .030$, $\eta_p^2 = .22$, $BF_{incl} = 2.96e+0$. As expected, overall modus ponens inferences were endorsed more strongly than modus tollens inferences. The main effect of mode was also significant, $F(1.37, 26.05) = 58.06$, $p < .001$, $\eta_p^2 = .75$, $BF_{incl} = 6.62e+12$. As expected, overall counterarguments reduced endorsement ratings. The interaction effect of inference and mode was significant, $F(2, 38) = 18.14$, $p < .001$, $\eta_p^2 = .49$, $BF_{incl} = 2.05e+3$. As expected, the interaction of inference and mode predicted endorsement ratings. Figure 3 shows the mean endorsement ratings of Experiment 1.

A 2 (inference: modus ponens vs. modus tollens) × 2 (mode: none vs. subjunctive) repeated-measures ANOVA revealed a significant main effect for inference, $F(1, 19) = 26.00$, $p < .001$, $\eta_p^2 = .58$, $BF_{incl} = 5.02e+3$, indicating that modus ponens inferences were endorsed more strongly than modus tollens inferences. The main effect of mode was also significant, $F(1, 19) = 53.71$, $p < .001$, $\eta_p^2 = .74$, $BF_{incl} = 6.43e+8$, indicating that subjunctive counterarguments reduced endorsement ratings compared with no counterarguments. The interaction effect of inference and mode was not significant, $F(1, 19) = 0.49$, $p = .492$, $\eta_p^2 = .03$, $BF_{incl} = 0.34e+0$.

A 2 (inference: modus ponens vs. modus tollens) × 2 (mode: none vs. indicative)

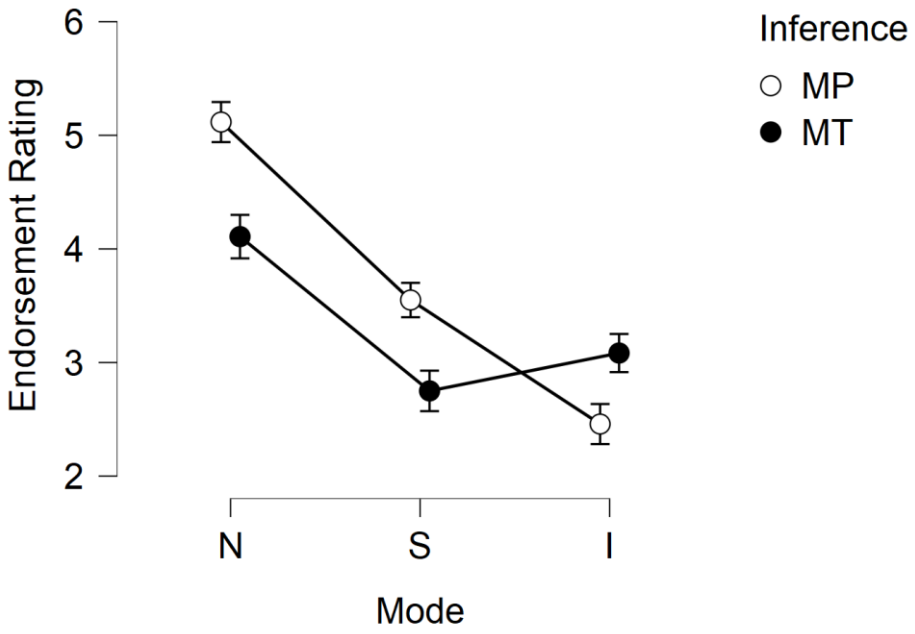

Figure 3. Mean endorsement ratings of Experiment 1 as a function of inference and mode. Inference: MP = Modus Ponens; MT = Modus Tollens. Mode: N = None; S = Subjunctive; I = Indicative. The scale ranges from 0 to 6. Error bars represent standard errors of the mean.

repeated-measures ANOVA showed no significant main effect of inference, $F(1, 19) = 1.49$, $p = .237$, $\eta_p^2 = .07$, $BF_{incl} = 0.34e+0$. The main effect of mode was significant, $F(1, 19) = 72.17$, $p < .001$, $\eta_p^2 = .79$, $BF_{incl} = 1.07e+10$, suggesting that indicative counterarguments reduced endorsement ratings compared with no counterarguments. However, this effect was qualified by mode, as the interaction effect of inference and mode was significant, $F(1, 19) = 23.71$, $p < .001$, $\eta_p^2 = .56$, $BF_{incl} = 5.42e+2$. Paired samples *t*-tests revealed that endorsement ratings were significantly higher for modus ponens inferences with no counterargument compared with modus ponens inferences with an indicative counterargument, $t(19) = 9.27$, $p < .001$, $d = 2.07$ (95% CI [1.28, 2.85]), $BF_{10} = 7.42e+5$. Endorsement ratings were significantly higher for modus tollens inferences with no counterargument compared with modus tollens inferences with an indicative counterargument, $t(19) = 3.93$, $p = .002$, $d = 0.88$ (95% CI [0.35, 1.39]), BF_{10}

= 4.01e+1. Endorsement ratings were significantly higher for modus ponens inferences with no counterargument compared with modus tollens inferences with no counterargument, $t(19) = 5.02$, $p < .001$, $d = 1.12$ (95% CI [0.55, 1.68]), $BF_{10} = 3.52e+2$. Endorsement ratings were significantly lower for modus ponens inferences with an indicative counterargument compared with modus tollens inferences with an indicative counterargument, $t(19) = -2.45$, $p = .024$, $d = -0.55$ (95% CI [-1.01, -0.07]), $BF_{10} = 2.48e+0$. Figure 4 and Figure 5 show the respective Bayesian sequential analyses of the post-hoc comparisons.

A 2 (inference: modus ponens vs. modus tollens) × 2 (mode: subjunctive vs. indicative) repeated-measures ANOVA showed no significant main effect of inference, $F(1, 19) = 0.15$, $p = .702$, $\eta_p^2 = .01$, $BF_{incl} = 0.25e+0$. The main effect of mode was significant, $F(1, 19) = 13.18$, $p = .002$, $\eta_p^2 = .41$, $BF_{incl} = 1.44e+0$. As predicted, this effect was qualified by inference, evidenced by a significant interaction effect of inference and mode, $F(1, 19) = 33.67$, $p < .001$, $\eta_p^2 = .64$, $BF_{incl} = 5.33e+2$. Paired samples *t*-tests revealed that endorsement ratings were significantly higher for modus ponens inferences with a subjunctive counterargument compared with modus ponens inferences with an indicative counterargument, $t(19) = 7.65$, $p < .001$, $d = 1.71$ (95% CI [1.01, 2.40]), $BF_{10} = 5.00e+4$. Endorsement ratings were lower for modus tollens inferences with a subjunctive counterargument compared with modus tollens inferences with an indicative counterargument, $t(19) = -1.88$, $p = .076$, $d = -0.42$ (95% CI [-0.87, 0.04]), $BF_{10} = 1.01e+0$; note, however, that this effect was only marginally significant. As hypothesized, endorsement ratings were significantly higher for modus ponens inferences with a subjunctive counterargument compared with modus tollens inferences with a subjunctive counterargument, $t(19) = 3.10$, $p = .018$, $d = 0.69$ (95% CI [0.20, 1.18]), $BF_{10} = 7.93e+0$. As hypothesized, endorsement ratings were indeed

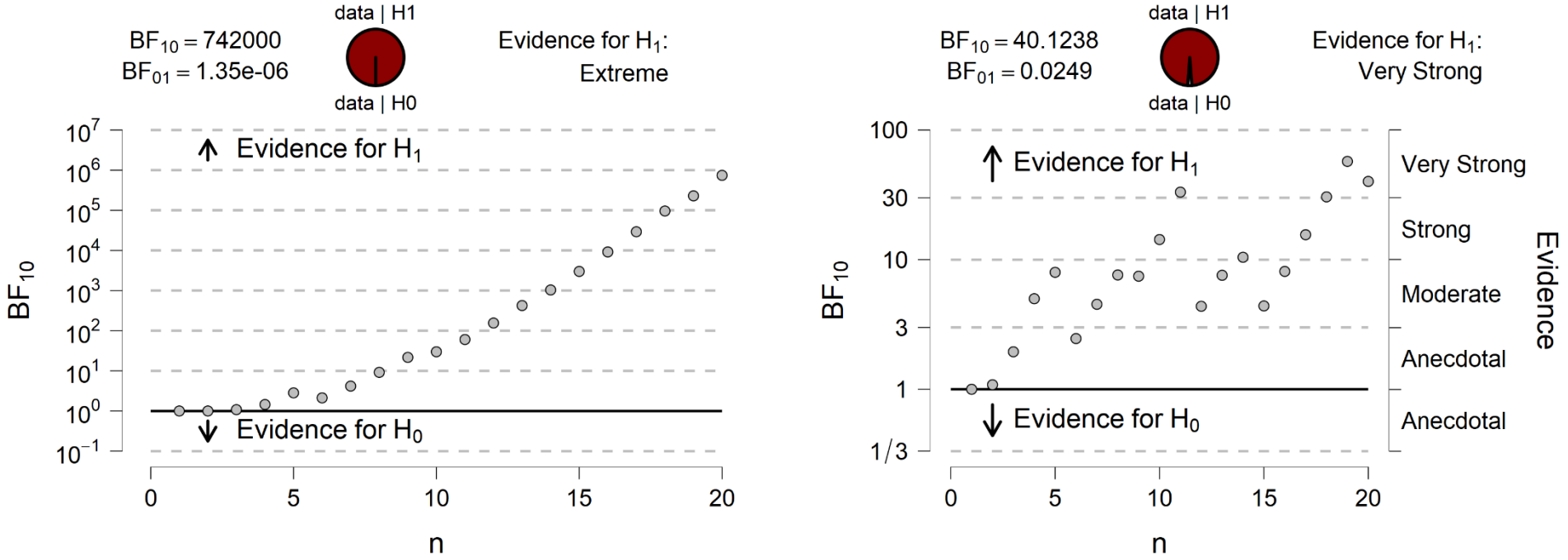

Figure 4. Bayesian sequential analysis of the endorsement ratings of Experiment 1. Left panel: difference between MP inferences with no counterargument versus MP inferences with indicative counterargument. Right panel: difference between MT inferences with no counterargument versus MT inferences with indicative counterargument.

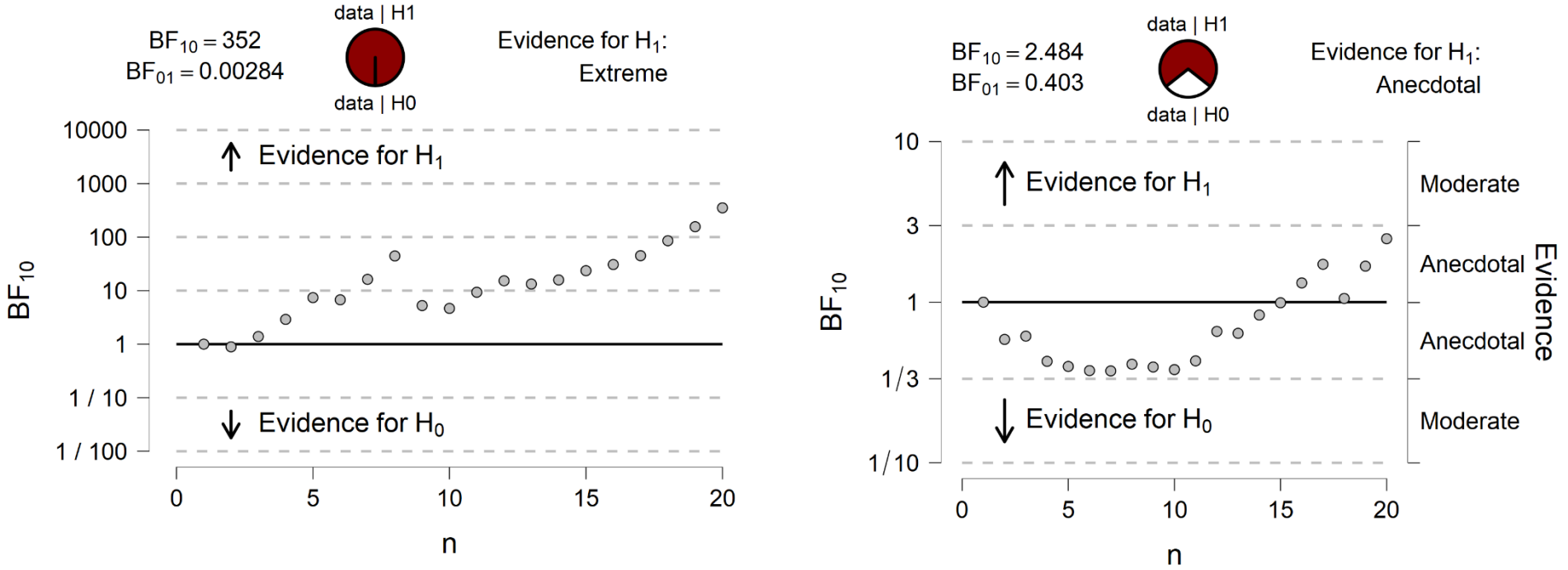

Figure 5. Bayesian sequential analysis of the endorsement ratings of Experiment 1. Left panel: difference between MP inferences with no counterargument versus MT inferences with no counterargument. Right panel: difference between MP inferences with indicative counterargument versus MT inferences with indicative counterargument.

significantly lower for modus ponens inferences with an indicative counterargument compared with modus tollens inferences with an indicative counterargument, $t(19) = -2.45$, $p = .048$, $d = -0.55$ (95% CI [-1.01, -0.07]), $BF_{10} = 2.48e+0$. Figure 6 and Figure 7 show the respective Bayesian sequential analyses of the post-hoc comparisons.

Response Times. The omnibus test, a 2 (inference: modus ponens vs. modus tollens) × 3 (mode: none vs. subjunctive vs. indicative) repeated-measures ANOVA, revealed a significant main effect for inference, $F(1, 19) = 6.08$, $p = .023$, $\eta_p^2 = .24$, $BF_{incl} = 5.25e+1$, indicating that overall responses times were lower for modus ponens inferences as opposed to modus tollens inferences. The main effect for mode was also significant, $F(1.52, 28.83) = 4.25$, $p = .033$, $\eta_p^2 = .18$, $BF_{incl} = 0.88e+0$, indicating that overall mode influenced response times. The interaction effect of inference and mode was not significant, $F(1.22, 23.25) = 0.83$, $p = .394$, $\eta_p^2 = .04$, $BF_{incl} = 0.21e+0$. Figure 8 shows the mean response times of Experiment 1.

3.1.4 Discussion

The aim of Experiment 1 was to study the interplay of inference type, counterarguments as well as counterarguments' linguistic mode and better comprehend how exactly this interplay affects conclusion endorsement. In line with our hypotheses, inference, mode, and the interaction between inference and mode predicted conclusion endorsement.

As expected, overall modus ponens inferences were endorsed more strongly than modus tollens inferences. This finding was not surprising because a large body of past research has consistently demonstrated that humans more frequently endorse modus ponens compared with modus tollens even though there is no logical reason to do so. A widely accepted explanation is that the modus tollens inference demands more cognitive resources than the modus ponens inference. The increased complexity of modus tollens in turn leads to more logical errors and unexpected results (Johnson-

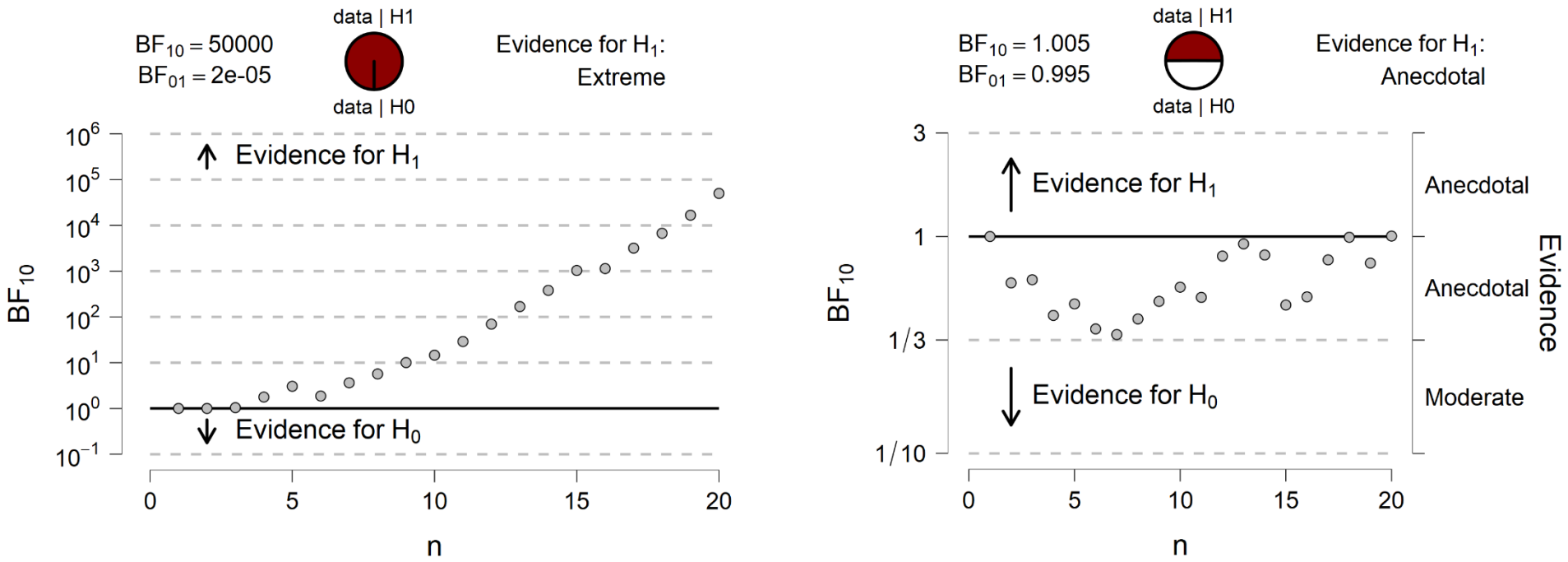

Figure 6. Bayesian sequential analysis of the endorsement ratings of Experiment 1. Left panel: difference between MP inferences with subjunctive counterargument versus MP inferences with indicative counterargument. Right panel: difference between MT inferences with subjunctive counterargument versus MT inferences with indicative counterargument.

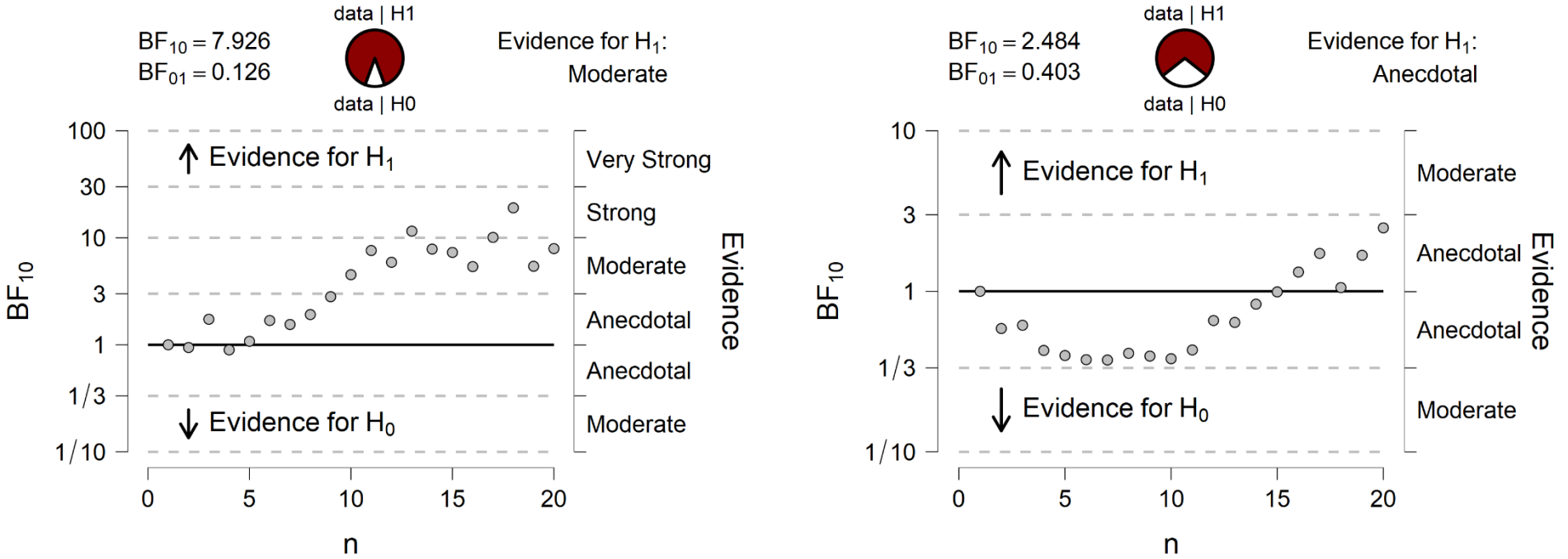

Figure 7. Bayesian sequential analysis of the endorsement ratings of Experiment 1. Left panel: difference between MP inferences with subjunctive counterargument versus MT inferences with subjunctive counterargument. Right panel: difference between MP inferences with indicative counterargument versus MT inferences with indicative counterargument.

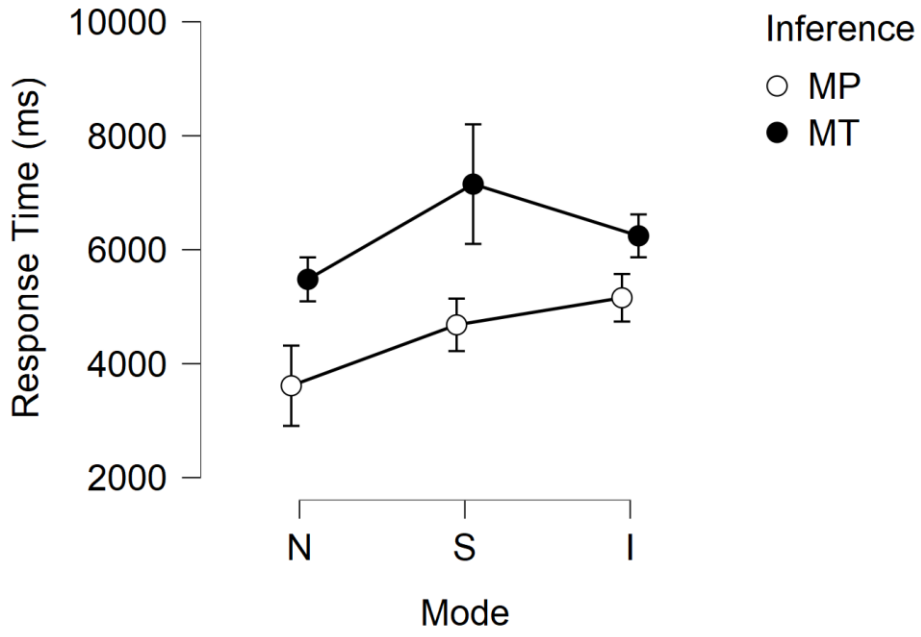

Figure 8. Mean response times of Experiment 1 as a function of inference and mode. Inference: MP = Modus Ponens; MT = Modus Tollens. Mode: N = None; S = Subjunctive; I = Indicative. Error bars represent standard errors of the mean.

Laird & Byrne, 1991; Singmann et al., 2014). A common interpretation of this processing disadvantage is informed by the fact that modus tollens requires the processing of negational statements, such as computing the negated consequent as well as actively negating the antecedent during the conclusion phase of the inference process. Reasoning theories typically state that negations demand more working memory capacity, thus putting more cognitive load onto the reasoning process (Braine & O'Brien, 1998; Johnson-Laird & Byrne, 1991; Oaksford et al., 2000). Consequently, endorsement accuracy decreases. The account to interpret the lower endorsement ratings for the modus tollens inferences in terms of the increased processing effort due to the encoding of negations is supported by the analysis of the collected response times. Participants needed significantly more time to respond to modus tollens inferences compared with modus ponens inferences. This finding suggests that modus tollens inferences indeed required more working memory resources to be solved. It is

worthwhile to also consider and discuss an alternative explanation for this effect that has been proposed by a few authors but did not receive as much attention. This alternative approach makes sense of the modus ponens and modus tollens disparity from a conversational perspective rather than from a mechanistic perspective. According to this intriguing approach, lower endorsement for modus tollens inferences occurs because reasoners are led to assume that the antecedent is the case although the consequent is not. According to the principle of relevance formulated by Sperber and Wilson (1995), people draw conclusions based on the messages they receive while assuming that the information transported by the message is relevant to them. Thereby, the message that a consequent is negated is only relevant if there are reasons to assume that an antecedent was actually given. Therefore, if reasoners think that some antecedent is the case and find out that the consequent is negated, they should implicitly consider that some exception occurred. Consequently, modus tollens inferences can be rejected. This way of responding is funnily referred to as the modus shmollens inference: Given the major premise “if p than q” and the minor premise “¬q” humans sometimes infer that “p” is the case regardless. Indeed, Bonnefon and Villejoubert (2007) showed that when the negation of a situation is represented in a statement uttered in the context of a conversation, a majority of reasoners endorse the modus shmollens inference, which makes no sense from a logical perspective but perfect sense from a conversational perspective.

Also in line with my a-priori theorizing is the empirically detected effect of mode. Overall, the presentation of a counterargument led to a decline of conclusion endorsement. This was true for both subjunctive and indicative counterarguments when compared against the condition in which no counterargument was presented. This finding supports the notion that conditional reasoning is not only deductive but also defeasible. Otherwise valid conclusions can be withdrawn in the light of new

evidence (Bonnefon & Vautier, 2010; Evans, 2002, 2012; Oaksford & Chater, 2009). Pollock (1987) argued that defeasibility is grounded in the human rational architecture. According to his theory, reasoning is guided by internalized rules that function as procedural knowledge. Defeasibility describes the idea that humans do not always conform to these rules, but this is because the production system that these rules constitute is embedded in a larger system that can override it. Therefore, what we should do is what the rules of our production system tell us to do, but this is not always what we actually do. Hence, just as in language use, we also have a competence versus performance distinction in reasoning. This form of defeasibility of inferences can be elicited via additional contextual information, which can be communicated by means of a counterargument (Byrne, 1991). Elqayam et al. (2015) contends that this capacity of the human mind to consider and integrate counterarguments is also a central feature of pragmatic reasoning. Pragmatic reasoning depends on encoding context such as background knowledge and new evidence and thus needs to take into account the possibility that formally correct conclusions may be altered when additional information is at hand. Again, the assessment of the corresponding response times revealed that the effect of counterarguments is sensible from a processing perspective, too. Overall, the presentation of counterarguments increased response times. Both subjunctive and indicative counterarguments increased response times when compared against the condition in which no counterargument was presented. I suppose that the additional encoding as well as integration of the counterargument into the inference process necessitates time, and this in turn directly manifested in the dependent measure for the response latencies.

As hypothesized, my findings revealed a significant interaction effect between inference and mode. A first implication one can draw from this finding is that conditional reasoning is affected by linguistic factors. Gazzo Castañeda and Knauff (2016b, 2019,

2021c) have already emphasized the importance of pragmatic factors for reasoning and provided first evidence to support their assertion. However, to date the pragmatic aspects of counterarguments and their impact on conditional reasoning remained uninvestiagted. Thus, this finding is the first to show that linguistic mode of counterarguments differentially affects conclusion endorsement for different inference types. As most reasoning takes place in semantically rich context, it is incontestable that pragmatics are crucial to understanding reasoning and that deductive reasoning requires a model of interpretation (Manktelow & Fairley, 2000). One building block of such an interpretational model is the notion that pragmatic factors such as the linguistic mode of a counterargument modulates the sufficiency and necessity of the causal link between the antecedent and the consequent of a conditional. In this way, a counterargument can create a disruption between a stated cause and a stated effect within a conditional. The extent to which this disruption moderates the final conclusion, in turn, is qualified by the inference type of the task as well as moderated by virtue of subtle linguistic variations in the respective counterarguments. Specifically, in accordance with my initial hypothesis concerning the interaction between inference and mode, I was able to predict conclusion endorsement. The predicted pattern of results could be observed. In particular, endorsement ratings were higher for modus ponens than for modus tollens inferences when a subjunctive counterargument was presented. In clear contrast, this pattern reversed in the other condition: Endorsement ratings were higher for modus tollens than for modus ponens inferences when an indicative counterargument was displayed. The Bayesian sequential analysis corroborated this finding. Note, however, that this analysis revealed that the evidence for a difference between modus ponens and modus tollens was more pronounced in the subjunctive condition than in the indicative condition. Subjective mode might have activated possibilities for exceptions to the rule (Johnson-Laird & Byrne, 2002; Quelhas

& Byrne, 2003). Possibilities were arguably not strong enough to trigger an active search process for further counterarguments, which resulted in higher endorsement for modus ponens compared with modus tollens because it is generally more easy to infer. However, indicative mode might have been represented as a factual exception to the rule, which evoked a search in semantic memory for more counterarguments in order to justify a rejection of the conclusion. I argue that this search process was inhibited for modus tollens because of the increased processing effort for this inference. Thus, less additional counterarguments were decoded, which eventually led to less exceptions to the rule and hence higher endorsement in comparison with modus ponens. I am aware that this is a highly mechanistic interpretation of the observed interaction pattern and thus requires further testing and the reproduction via alternative analytical strategies. I perform and report the necessary steps to validate this finding in the next chapters. Moreover, my algorithmic approach to account for the pattern of results is corroborated by previous work showing that reasoners indeed engage in a semantic search process, that counterargument search does not stop after the first counterargument is presented or retrieved, and that every additional counterargument weakens the certainty regarding the necessity of the causal link between antecedent and consequent (De Neys et al., 2003a, 2003b; Politzer & Bourmaud, 2002). An alternative interpretation to account for the interaction pattern stresses the concept of relative salience. Chan and Chua (1994) found that the relative salience of a counterargument, that is the associative strength between the conditional and a counterargument, influences endorsement ratings. According to this idea, indicative counterarguments may exhibit a higher relative salience in the case of modus ponens. Modus ponens inferences are far more common in everyday life and decision making. Thus, strong indicative counterarguments might display higher relative salience because more of them are stored in memory and hence they are more easily

accessible. In contrast, modus tollens inferences are rarer in everyday reasoning; hence, less counterarguments are available and consequently relative salience decreases. This principle should only apply when indicative counterarguments are given because subjunctive counterarguments are epistemically too weak to induce the alleged search process. A recent finding from Skovgaard-Olsen and Collins (2021) suggests that indicative premises convey neutrality, whereas subjunctive premises may convey falsity as conversational implicature. My findings extend this insight because they indicate that the conversational implicature conveyed by the linguistic mode of the counterargument has differential effects on conclusion endorsement depending on the underlying inference type.

Please note that the analysis of the secondary measure indicated no significant interaction of inference and mode for the response times. This finding is puzzling insofar as it seems to attenuate or speak against the process-level explanation of the interaction effect in the endorsement ratings. In case both the significant interaction in endorsement ratings as well as the non-significant interaction in response times will be replicated in the following experiments and reproduced in the additional follow-up analyses, then a functional interpretation of the findings is presumably a better suited and more parsimonious account to explain the pattern of results because it relies on a computational level of cognitive analysis, which unlike an algorithmic level does not necessitate to make assumptions about the internal processing mechanisms driving the effects. Yet, it is still conceivable to interpret findings of endorsement ratings and response times discretely. Endorsement ratings only reflect a behavioral response, whereas response times cover the latency between a stimulus onset, stimulus perception, cognitive processing, action planning, and the final response. Hence, response times can incorporate and reflect various stages of processing from stimulus input to response output. It is therefore problematic to assume a one-to-one

correspondence between endorsement ratings and response times. This suggests that at best the exploratory analysis of response times can be used to complement the findings from the endorsement ratings. However, using response times in order to invalidate behavioral results obtained via endorsement ratings is unwarranted. Thus, I decide to interpret the response times with caution and to put the emphasis on the endorsement ratings as the central dependent measure.

3.2 Experiment 2

Experiment 2 served as a replication study of the previous experiment. Replication constitutes a central scientific method that serves the purpose of corroborating the validity of an original research finding. Aside from other questionable research practices like HARKing, cherry-picking, *p*-hacking, fishing expeditions, data dredging, and publication bias (e.g., see Ioannidis, 2005; Kerr, 1998; Vul et al., 2009), the lack of a replication study can inflate type I error rates and thus produce false positives (John et al., 2012; Nosek et al., 2022; Shrout & Rodgers, 2018; Simmons et al., 2011; Simons, 2014; Simonsohn, 2015). Even decades before it became public that psychology as a field suffered from a replication crisis (Anderson & Maxwell, 2017; Giner-Sorolla, 2019; Hughes, 2018), among other epistemological issues warnings have been spelled out that replicability of study findings might be lower than initially expected (Button et al., 2013; Cohen, 1973, 1992, 1994; Greenwald, 1975; Meehl, 1978; Rosenthal, 1979; Sedlmeier & Gigerenzer, 1992; Sterling, 1959; Szucs & Ioannidis, 2017). In 2005, Ioannidis argued that in a field where statistical significance is declared an implicit presupposition for publication, “[t]here is increasing concern that most current published research findings are false” (p. 696). In fact, a large-scale research collaboration that has set out to replicate 100 original studies, which were published in three major high-impact journals, has shown replication effects only amounted to half the magnitude of the original studies’ effects. Of the original studies,

97% displayed statistically significant effects, whereas merely 36% of the replication studies displayed statistically significant effects (Open Science Collaboration, 2015). Further work also revealed the detrimental ramifications of a lack of replication for the credibility of research outputs (Bakker et al., 2012; Simmons et al., 2011; Wagenmakers et al., 2011). It has been shown that especially the life sciences (i.e., disciplines such as medicine, psychology, genetics, biology) are confronted with and affected by a failure to replicate previous findings (Schooler, 2014). All the more it becomes increasingly imperative to regard replication studies as a sine qua non condition to render novel empirical findings credible (Schmidt, 2009). As Simons (2014) stated, “if an effect is real and robust, any competent researcher should be able to obtain it when using the same procedures with adequate statistical power” (p. 76). While the proposed ways on how to overcome the replication crisis are manifold and far from consensual (Mayo, 2021), the fact that replication increases confidence in original study results is considered common ground (Amrhein et al., 2019). An additional advantage of running rigorously operationalized replication studies is that they counteract selection bias and measurement errors of individual studies (Loken & Gelman, 2017). While various causes contributing to the replication crisis have been identified by now (Lewandowsky & Oberauer, 2020), it is clear that the rigorous implementation of replication studies can be a way forward to help solve it (Schooler, 2014).

Consequently, I decided to run a replication in order to validate the findings of the previous experiment. Experiment 2 was a direct replication study, meaning that as many variables and conditions as possible were held constant in comparison to Experiment 1. However, one aspect of the study that necessarily had to be adapted was the language of the materials because the study was conducted at an Italian university. Hence, the method of back-translation (Brislin, 1970) was applied in order

to assure a valid and reliable translation of the study materials. In this sense, one might argue that another element of the study context has changed as well, namely the population from which the participants for Experiment 2 were sampled. Indeed, the specific language-culture group an experiment is conducted in can—if certain conditions are met—exert an impact on study outcomes. The field of cultural psychology has generated impressive knowledge and continuously provides intriguing insights into the moderating conditions that qualify and the mediating mechanisms that produce the similarities as well as the differences when comparing different language-culture groups (for a review, see Oyserman, 2017). In general, there appear to exist two major schools dominating the debate on whether or not language and cultural context influence fundamental cognitive processes and representations and hence may impact cognitively driven outcome variables such as conclusions, judgments, and decisions. These schools may be termed the “language affects cognition” school on the one hand, and the “language does not affect cognition” school on the other hand. The former school was largely informed by the Sapir-Whorf hypothesis, also known as the linguistic relativity hypothesis (Kay & Kempton, 1984; Mandelbaum, 1951; Sapir, 1921; Whorf, 1956). The Sapir-Whorf hypothesis posits that language as a human prerogative in general and specific languages in particular determine cognition by virtue of reorganizing and restructuring its underlying processes, even in domains such as spatial reasoning (e.g., Levinson et al., 2002) that falsely have been considered natural and universal. Instead, it is argued that thinking is language-dependent in a plethora of important domains—therefore, the influence of language on restructuring thought explains many of the special properties of human thinking and reasoning (e.g., Bowerman & Levinson, 2001; Everett, 2005; Frank et al., 2008; Lucy, 1992a, 1992b; Spelke & Tsivkin, 2001). In stark contrast, the latter school was largely informed by the theory of universal grammar (Chomsky, 1965). The theory of universal grammar posits

that language constitutes a functional system for an innate language of thought (Fodor, 1975). Hence, language either explicitly reflects an antecedently available module of universal concepts, or it is based on a rich core set of natural concepts that build a universal conceptual foundation (Pinker, 1994), which is ultimately independent of culture-specific practices and the pragmatic peculiarities between single languages (e.g., Knauff & Ragni, 2011; Li & Gleitman, 2002). While the main objective of Experiment 2 was to examine whether the obtained results from the previous experiment replicate *ceteris paribus*, the new study context allowed me to compare the findings between the two language-culture groups. Therefore, the comparison between the results from Experiments 1 and 2 will also provide a small contribution for the ongoing debate outlined above.

3.2.1 Hypothesis

In general, I tested whether the results from the previous experiment are replicable in Experiment 2. I expected that (1) overall, modus ponens inferences display higher endorsement ratings than modus tollens inferences, (2) overall, counterarguments reduce endorsement ratings, and (3) the interaction of inference and mode predicts endorsement ratings. Specifically, I hypothesized that while for subjunctive mode endorsement ratings are higher for modus ponens inferences as opposed to modus tollens inferences, this relationship reverses for indicative mode. For indicative mode, endorsement ratings are lower for modus ponens inferences as opposed to modus tollens inferences. As for response times, I based my predictions based on the exploratory response time findings from Experiment 1. I hypothesized that response times are overall lower for modus ponens inferences than for modus tollens inferences. I further hypothesized that overall counterarguments increase response times. In line with the respective finding from Experiment 1, I assumed to

observe a null effect for the interaction between inference and mode in the response times.

3.2.2 Method

This section gives a detailed description of the sample characteristics, study design, materials, and experimental procedures of Experiment 2.

Participants. Given that Experiment 2 served as a direct replication study of the previous experiment, I aimed to keep study conditions as constant as possible in order to minimize a potential confounding via covariates. Therefore, I also kept sample size constant. Hence, I again collected data from $N = 20$ participants. Accordingly, the a-priori power analysis via G*Power (Faul et al., 2007, 2009) estimated 19 participants to be sufficient to reliably detect a medium effect size of Cohen's $f = .25$ (Cohen, 1988; Ellis, 2010), based on a type I error of $\alpha = .05$ and a power of $1 - \beta = .80$. The respective sensitivity analysis revealed that 20 participants are sufficient to reliably observe a minimum detectable effect (MDE) of Cohen's $f = .24$, given a type I error of $\alpha = .05$ and assuming a power of $1 - \beta = .80$. The final sample consisted of Italian university students of different majors who were between 19 and 29 years old ($M = 23.65$, $SD = 2.64$; 6 male). The first language of all participants was Italian. No participant had prior expertise in formal logic. Participants were recruited by sending a circular email via the email server of the Sapienza University of Rome, Italy. They were paid or received course credit for their participation.

Design. In line with the previous experiment, Experiment 2 followed a 2 (inference: modus ponens vs. modus tollens) × 3 (mode: none vs. subjunctive vs. indicative) within-participants design. Dependent measures were endorsement rating and response time.

Materials. The study materials were identical to Experiment 1 with the exception that all instructions, stimuli, and response formats were translated into Italian language

beforehand. I used back-translation as a translation method, which was first introduced by Brislin (1970) for cross-cultural research. It constitutes the methodological gold standard to ensure high-quality translations and is the most commonly applied procedure to minimize various sources of bias and to maximize the validity of the linguistic adaptations during the translation process (e.g., Tyupa, 2011; Van de Vijver & Leung, 2011). For this purpose, two German-Italian bilinguals (henceforth: T1 and T2) assisted as translators of the study materials. The whole translation procedure comprised four steps: In step 1, T1 created a forward-translation from the source language (i.e., German) to the target language (i.e., Italian). In step 2, T2 independently created a back-translation from the target language to the source language. In step 3, I as the investigator received the back-translated German version and compared it with the original German version in order to check for equivalence between the original and the back-translation, and to identify translation problems. In step 4, T1, T2, and myself discussed remaining minor issues concerning small linguistic deviations, as well as queries referring to subtle cultural adaptations whenever a literal translation would have produced a culturally inequivalent item. Based on these considerations, final revisions were made in the translated Italian version. Thereby, a high degree of syntactic, semantic, and pragmatic equivalence between the original German version and the translated Italian version of the study materials was assured. All instructions (Appendix B1), stimuli (Appendix B2), and response formats (Appendix B3) of Experiment 2 are appended.

Procedure. The experiment was implemented and conducted with the program OpenSesame (Mathôt et al., 2012). The procedure of Experiment 2 was identical to the procedure of the previous experiment, with the exception that all instructions, stimuli, and response formats were presented in the translated Italian version of the study materials.

3.2.3 Results

All participants met the inclusion criteria (i.e., Italian as first language, no expertise in formal logic) and completed the experiment without the occurrence of technical errors. Thus, no data had to be excluded. The statistical software packages, the data pre-processing (i.e., structuring, cleaning, transforming, aggregating), and the statistical tests were identical to the previous experiment.

Endorsement Ratings. The omnibus test, a 2 (inference: modus ponens vs. modus tollens) × 3 (mode: none vs. subjunctive vs. indicative) repeated-measures ANOVA, indicated that the main effect of inference did not reach significance, $F(1, 19) = 2.84$, $p = .108$, $\eta_p^2 = .13$, $BF_{incl} = 0.45e+0$. That is, overall modus ponens inferences and modus tollens inferences did not significantly differ in endorsement ratings. The main effect of mode was significant, $F(1.49, 28.32) = 42.42$, $p < .001$, $\eta_p^2 = .69$, $BF_{incl} = 3.15e+15$. As expected, overall counterarguments reduced endorsement ratings. The interaction effect of inference and mode was also significant, $F(2, 38) = 13.65$, $p < .001$, $\eta_p^2 = .42$, $BF_{incl} = 2.25e+1$. As expected, the interaction of inference and mode predicted endorsement ratings. Figure 9 shows the mean endorsement ratings of Experiment 2.

A 2 (inference: modus ponens vs. modus tollens) × 2 (mode: none vs. subjunctive) repeated-measures ANOVA showed a significant main effect for inference, $F(1, 19) = 13.85$, $p = .001$, $\eta_p^2 = .42$, $BF_{incl} = 2.94e+1$, indicating that modus ponens inferences were endorsed more strongly than modus tollens inferences. The main effect of mode was also significant, $F(1, 19) = 59.15$, $p < .001$, $\eta_p^2 = .76$, $BF_{incl} = 4.93e+11$, indicating that subjunctive counterarguments reduced endorsement ratings compared with no counterarguments. The interaction effect of inference and mode was not significant, $F(1, 19) = 0.03$, $p = .873$, $\eta_p^2 = .00$, $BF_{incl} = 0.31e+0$.

A 2 (inference: modus ponens vs. modus tollens) × 2 (mode: none vs. indicative)

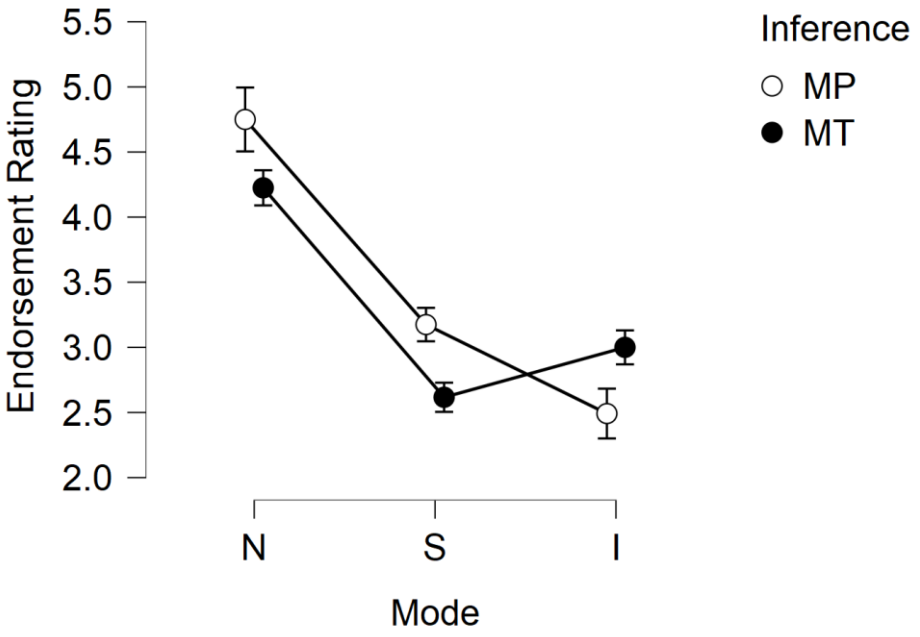

Figure 9. Mean endorsement ratings of Experiment 2 as a function of inference and mode. Inference: MP = Modus Ponens; MT = Modus Tollens. Mode: N = None; S = Subjunctive; I = Indicative. The scale ranges from 0 to 6. Error bars represent standard errors of the mean.

repeated-measures ANOVA showed no significant main effect of inference, $F(1, 19) = 0.00$, $p = .953$, $\eta_p^2 = .00$, $BF_{incl} = 0.23e+0$. The main effect of mode was significant, $F(1, 19) = 45.48$, $p < .001$, $\eta_p^2 = .71$, $BF_{incl} = 1.78e+10$, suggesting that indicative counterarguments reduced endorsement ratings compared with no counterarguments. However, this effect was qualified by mode, as the interaction effect of inference and mode was significant, $F(1, 19) = 16.14$, $p < .001$, $\eta_p^2 = .46$, $BF_{incl} = 6.39e+0$. Paired samples *t*-tests revealed that endorsement ratings were significantly higher for modus ponens inferences with no counterargument compared with modus ponens inferences with an indicative counterargument, $t(19) = 6.42$, $p < .001$, $d = 1.44$ (95% CI [0.80, 2.06]), $BF_{10} = 5.36e+3$. Endorsement ratings were significantly higher for modus tollens inferences with no counterargument compared with modus tollens inferences with an indicative counterargument, $t(19) = 5.94$, $p < .001$, $d = 1.33$ (95% CI [0.71, 1.93]), BF_{10}

= 2.15e+3. Endorsement ratings were significantly higher for modus ponens inferences with no counterargument compared with modus tollens inferences with no counterargument, $t(19) = 2.53$, $p = .021$, $d = 0.57$ (95% CI [0.09, 1.03]), $BF_{10} = 2.83e+0$. Endorsement ratings were significantly lower for modus ponens inferences with an indicative counterargument compared with modus tollens inferences with an indicative counterargument, $t(19) = -2.97$, $p = .016$, $d = -0.66$ (95% CI [-1.14, -0.17]), $BF_{10} = 6.25e+0$. Figure 10 and Figure 11 show the respective Bayesian sequential analyses of the post-hoc comparisons.

A 2 (inference: modus ponens vs. modus tollens) × 2 (mode: subjunctive vs. indicative) repeated-measures ANOVA showed no significant main effect of inference, $F(1, 19) = 0.06$, $p = .817$, $\eta_p^2 = .00$, $BF_{incl} = 0.24e+0$. The main effect of mode was also not significant, $F(1, 19) = 1.01$, $p = .327$, $\eta_p^2 = .05$, $BF_{incl} = 0.37e+0$. As predicted, I found a significant interaction effect of inference and mode, $F(1, 19) = 21.30$, $p < .001$, $\eta_p^2 = .53$, $BF_{incl} = 3.60e+2$. Paird samples *t*-tests revealed that endorsement ratings were significantly higher for modus ponens inferences with a subjunctive counterargument compared with modus ponens inferences with an indicative counterargument, $t(19) = 2.93$, $p = .024$, $d = 0.66$ (95% CI [0.16, 1.13]), $BF_{10} = 5.82e+0$. Endorsement ratings were significantly lower for modus tollens inferences with a subjunctive counterargument compared with modus tollens inferences with an indicative counterargument, $t(19) = -2.96$, $p = .024$, $d = -0.66$ (95% CI [-1.14, -0.17]), $BF_{10} = 6.09e+0$. As hypothesized, endorsement ratings were significantly higher for modus ponens inferences with a subjunctive counterargument compared with modus tollens inferences with a subjunctive counterargument, $t(19) = 3.93$, $p = .004$, $d = 0.88$ (95% CI [0.35, 1.39]), $BF_{10} = 3.99e+1$. As hypothesized, endorsement ratings were significantly lower for modus ponens inferences with an indicative counterargument compared with modus tollens inferences with an indicative counterargument, $t(19) = -$

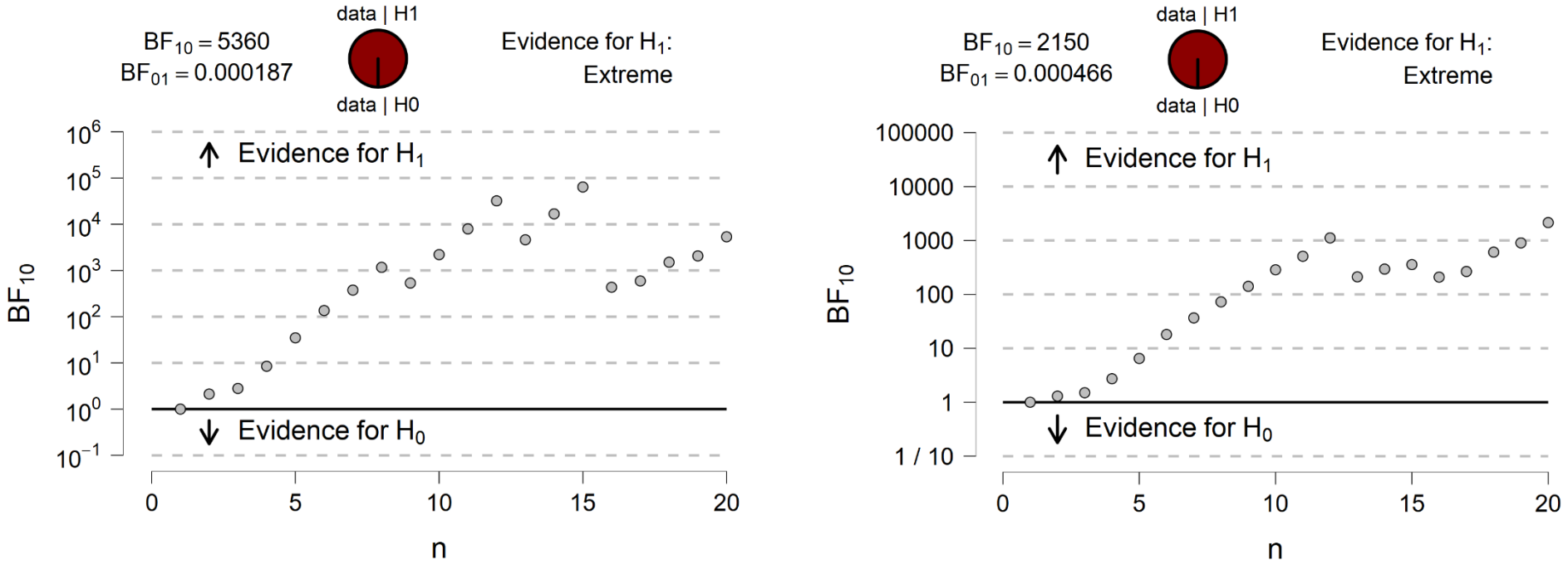

Figure 10. Bayesian sequential analysis of the endorsement ratings of Experiment 2. Left panel: difference between MP inferences with no counterargument versus MP inferences with indicative counterargument. Right panel: difference between MT inferences with no counterargument versus MT inferences with indicative counterargument.

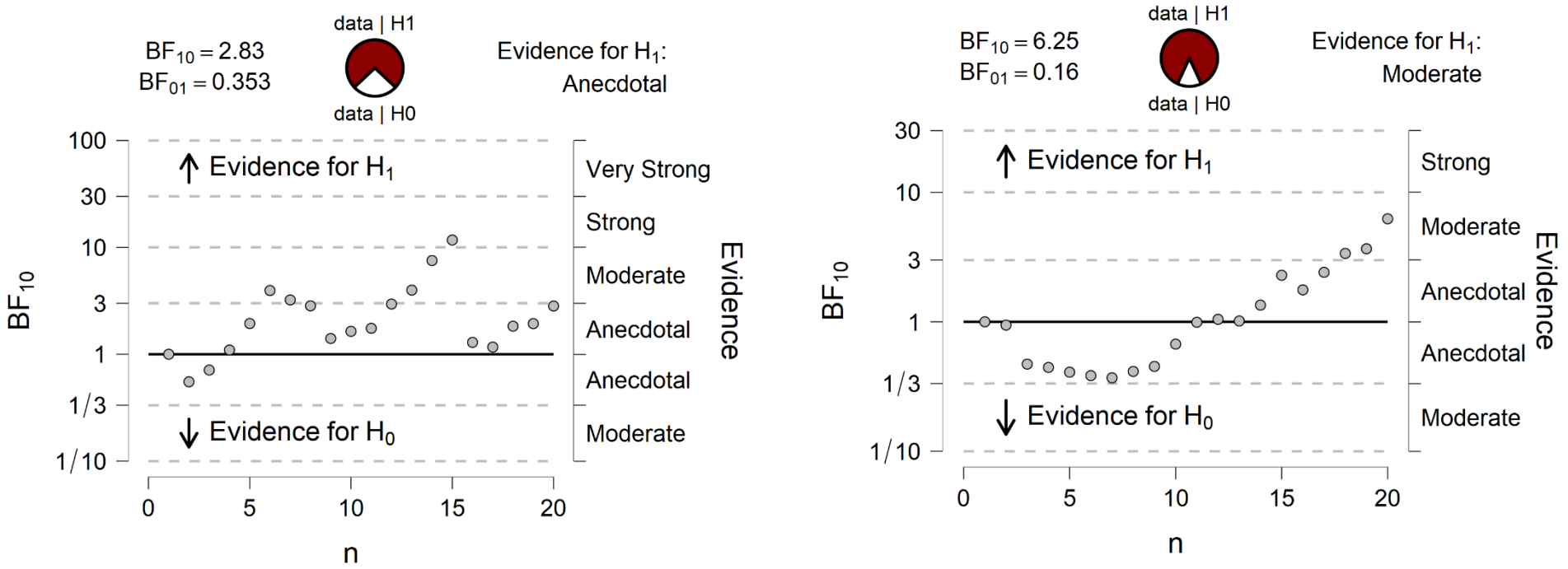

Figure 11. Bayesian sequential analysis of the endorsement ratings of Experiment 2. Left panel: difference between MP inferences with no counterargument versus MT inferences with no counterargument. Right panel: difference between MP inferences with indicative counterargument versus MT inferences with indicative counterargument.

2.97, p = .024, d = -0.66 (95% CI [-1.14, -0.17]), BF_{10} = 6.25e+0. Figure 12 and Figure 13 show the respective Bayesian sequential analyses of the post-hoc comparisons.

Response Times. The omnibus test, a 2 (inference: modus ponens vs. modus tollens) × 3 (mode: none vs. subjunctive vs. indicative) repeated-measures ANOVA, revealed a significant main effect for inference, F(1, 19) = 19.03, p < .001, η_p^2 =.50, BF_{incl} = 5.17e+0, indicating that overall responses times were lower for modus ponens inferences as opposed to modus tollens inferences. The main effect for mode was marginally significant, F(2, 38) = 2.73, p = .078, η_p^2 =.13, BF_{incl} = 2.53e+0, suggesting that overall mode influenced response times. The interaction effect of inference and mode was not significant, F(1.50, 28.50) = 0.71, p = .462, η_p^2 =.04, BF_{incl} = 0.21e+0. Figure 14 shows the mean response times of Experiment 2.

3.2.4 Discussion

The aim of Experiment 2 was to conduct a replication study that examines the role of inference type, counterarguments, and counterarguments' linguistic mode on the conclusions inferred when engaging in conditional reasoning. I will first discuss the results of Experiment 2 with a particular focus on their relation to the results of Experiment 1. Then, I will provide a critical review of the employed method of back-translation and elucidate strengths and pitfalls of this method. I also wish to reflect on the nature of the current replication study, compare definitions and categories of replication studies, and demarcate the present replication from other forms of replication studies. Eventually, I will briefly discuss what my findings mean for the debate between proponents of linguistic relativity and proponents of universal grammar.

Contrary to my expectation, the omnibus test showed no main effect for inference. However, I wish to emphasize that this null finding should be interpreted very cautiously for several reasons. First, the significance level of the effect did reach

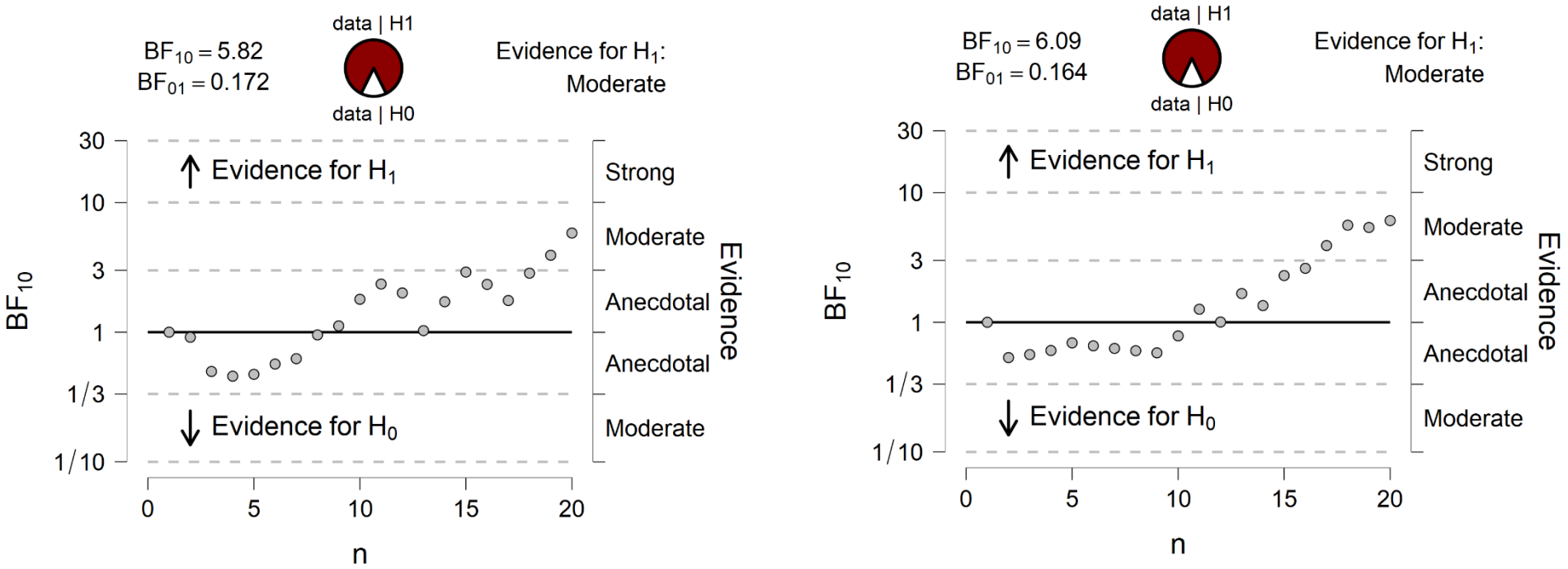

Figure 12. Bayesian sequential analysis of the endorsement ratings of Experiment 2. Left panel: difference between MP inferences with subjunctive counterargument versus MP inferences with indicative counterargument. Right panel: difference between MT inferences with subjunctive counterargument versus MT inferences with indicative counterargument.

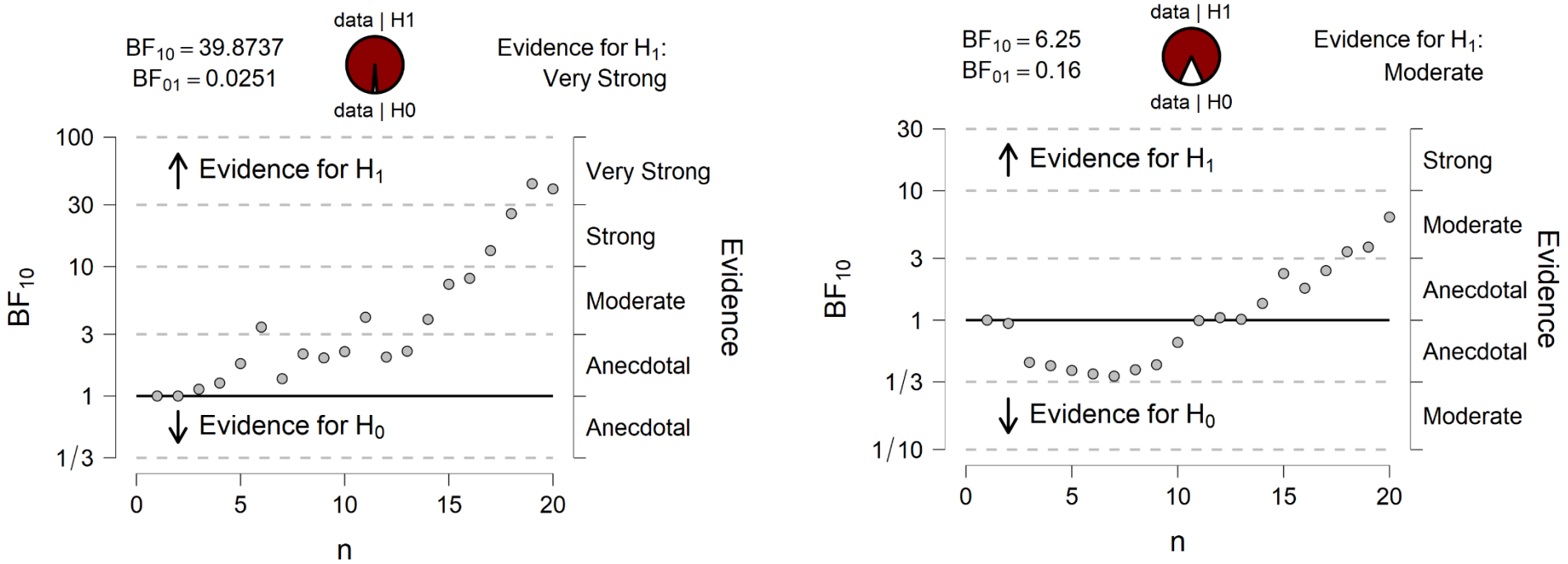

Figure 13. Bayesian sequential analysis of the endorsement ratings of Experiment 2. Left panel: difference between MP inferences with subjunctive counterargument versus MT inferences with subjunctive counterargument. Right panel: difference between MP inferences with indicative counterargument versus MT inferences with indicative counterargument.

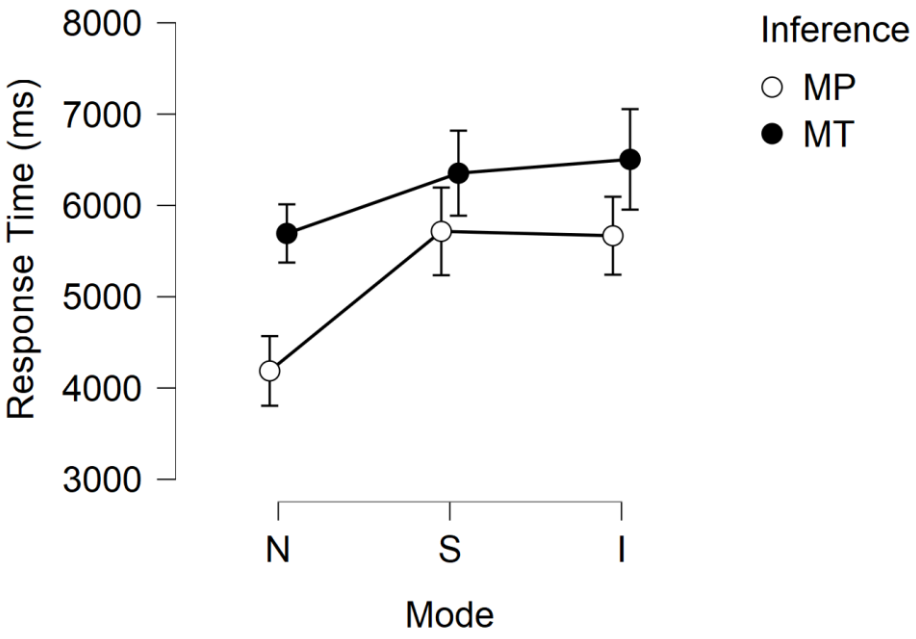

Figure 14. Mean response times of Experiment 2 as a function of inference and mode. Inference: MP = Modus Ponens; MT = Modus Tollens. Mode: N = None; S = Subjunctive; I = Indicative. Error bars represent standard errors of the mean.

the edge of marginal significance with p = .108. Besides, in recent years it became increasingly acknowledged that the p-value should not be the only criterion to decide whether or not a hypothesis is rejected. Instead, other measures such as effect size estimates and confidence intervals should be considered (Amrhein et al., 2019; Cohen, 1994; Colquhoun, 2014, 2017; Cumming, 2008, 2014; Cumming & Calin-Jageman, 2016; Halsey, 2019; Halsey et al., 2015; Huber, 2016; Lazzeroni et al., 2016; Nuzzo, 2014; Wasserstein & Lazar, 2016). Indeed, the respective effect size for the main effect of inference amounts to η_p^2 = .13. The widely agreed upon conventions for ANOVA effect sizes are η_p^2 = .01 for a small effect, η_p^2 = .06 for a moderate effect, and η_p^2 = .14 for a large effect (Cohen, 1988; Ellis, 2010). This means that inference displays a moderate effect size; it is even adjacent to the threshold of a large effect. Thus, I argue that the main effect of inference is definitely of practical significance (despite its lack of

statistical significance given the *p*-value as a conservative benchmark). Second, the inspection of the pattern of results conveys that the results of Experiment 2 follow the same pattern as those of Experiment 1. Albeit the significance levels of the main effect of inference differ, the pattern of effects conveys close similarity and suggests that the actually observed descriptive differences are only minor ones. Third, it is difficult to interpret main effects when the interaction effect is significant in the first place, which was the case here. This is another reason that relativizes this finding. And fourth, the respective hypothesis regarding the response times was confirmed: Response times for modus ponens inferences were lower than response times for modus tollens inferences. This corroborates the claim of modus tollens inferences being generally more challenging inferences to draw than modus ponens inferences (e.g., Singmann et al., 2014).

As predicted, the main effect of mode was significant. Overall, the presentation of a counterargument decreased conclusion endorsement. This was again true for both subjunctive and indicative counterarguments when compared against the condition in which no counterargument was presented. This replicated finding speaks for the notion that conditional reasoning is affected by more than formal logic (e.g., Evans, 2012) and that counterarguments exert a direct influence in the inferences we draw. Correspondingly, response times increased when counterarguments were presented. Again, both subjunctive and indicative counterarguments increased response times when compared against the condition in which no counterargument was presented.

As hypothesized, and most importantly, the interaction effect between inference and mode was replicated in Experiment 2. The predicted pattern of results could be observed a second time. Endorsement ratings were higher for modus ponens than for modus tollens inferences when a subjunctive counterargument was presented, whereas endorsement ratings were higher for modus tollens than for modus ponens

inferences when an indicative counterargument was shown. This pattern was further corroborated by the Bayesian sequential analysis. Hence, the successful replication increases the confidence in the validity of the findings in general and the exact pattern of the interaction between inference and mode in particular. The fact that the relationship of subjunctive versus indicative counterarguments reverses when switching from modus ponens inference to modus tollens inference highlights the complex ramifications on reasoning that can be invoked by virtue of slight linguistic variations (e.g., Manktelow & Fairley, 2000; Manktelow et al., 2000). These variations appear to evoke changes in the representations of the pragmatic meaning of the conditionals, which ultimately manifest themselves in a pattern reversal of the dependent variable. However, I wish to point out that while I established a novel effect in Experiment 1, and replicated it in Experiment 2, starting to examine its moderators and boundary conditions is still pending. This task will be addressed in Experiment 3.

This next paragraph is added to discuss the advantages and limitations of the applied method of back-translation (Brislin, 1970, 1986). The central purpose of back-translation is to provide quality control during the translation of study materials since it assures a high degree of equivalence between the original materials and their translations (Behr, 2017). Although back-translation can vary in its concrete implementation, the general algorithm is the following: forward translation, back-translation, back translation review and discussion, and finalization. The third step is arguably the one that requires very careful revisions and good communication between the principal investigator, T1, and T2. It is the step that facilitates the fine-tuning of the translation (Tyupa, 2011). I suppose that the fact that we carried out all steps diligently (as well as in close cooperation between myself, T1, and T2 for the review and discussion procedure) has contributed to the success of the present replication. A major advantage of the back-translation technique, compared with other available

methods (e.g., Carroll, 1966; Miller & Beebe-Center, 1956), is that the principal investigator does not necessarily need to understand or speak the target language. Furthermore, back-translation does not require the construction of test items or scoring tests, no special equipment is needed, no technical experts as adjuncts are needed, the relative costs are low, and no test subjects are required (Sinaiko & Brislin, 1973). However, a comprehensive evaluation of the back-translation method must also involve critical perspectives that illuminate the weaknesses of this approach. Sinaiko and Brislin (1973) argued that one disadvantage of back-translation lies in the fact that any mistake can be either due to T1 or T2. Thus, it is not clear whether errors and inconsistencies occur because a good translation was misinterpreted by an incompetent T2, or because a T2 works with a translation that is flawed due to an incompetent T1. Put differently, discrepancies between the original version and the back-translated version can be due to errors in the actual translation but they can also be due to errors in the back-translation—or both (Harkness, 2003; Harkness et al., 2004). Moreover, for the sake of cultural adaptation it is sometimes necessary to intentionally modify items if this is the only means to create a valid measure for a new language-culture group (Behr, 2017). An additional tool to help balance cultural variations between original and translation is called decentering (Werner & Campbell, 1970). Here, the aim is not to translate one version of materials into another language with minimum change. Rather, decentering refers to a process by which the materials are adapted in a way that allows a smooth and naturally sounding version in the target language (Brislin, 1976). The study materials are not centered around one single language or culture. Rather, both the idiosyncratic features of original language and target language co-determine the study materials' final version. Hence, the translation quality of the back-translation technique can be improved when it is complemented with decentering. Another point worth mentioning concerns the existence of alternative

translation methods. The most important alternatives to back-translation are knowledge testing and performance testing. For further information about these alternatives and a comparison between the three methods, I refer the interested reader to Sinaiko and Brislin (1973).

A conceptual point of discussion concerns the different forms of replication studies. A widely accepted taxonomy for the categorization of different forms of replications is the divide between direct replication, systematic replication, and conceptual replication. A direct (or exact) replication is a new study employing the same study population, materials, design, measures, and procedure as the original study (Brandt et al., 2014). A systematic replication is basically a direct replication in which only minor ancillary features, which are assumed to be negligible, deviate from the original study. A conceptual replication is intentionally different from the original study. It is designed to assess the generalizability and the veracity of a finding. For example, it may comprise a similar but not identical manipulation, or samples from a distinctly different population or era (Fabrigar & Wegener, 2016; Ledgerwood et al., 2017). It is not trivial to say which of either categories Experiment 2 falls into. *Ceteris paribus*, the only thing that changed in comparison to Experiment 1 was the study population. Indeed, this is a different study population, which would suggest that Experiment 2 was a conceptual replication. However, are Italian university students really a distinctly different study population from German university students? I doubt it. The highly similar findings, both with respect to the endorsement ratings as well as the response times, speak against it. On the other hand, one can also not make the claim that it was a strictly identical direct replication because the study materials were translated. Maybe the most appropriate label for Experiment 2 would be that of a systematic replication, even though I am not completely content with this categorization either. After all, one should bear in mind that these categories are arbitrary and

artificial. The distinctions between direct, systematic, and conceptual replication represent a continuum (Shrout & Rodgers, 2018). All types of replications are informative—one just needs to be clear about the goal of the replication and the context to which one intends to generalize in advance. For example, while a direct replication shows that an effect is stable under certain, precisely specified conditions, a conceptual replication provides evidence that an effect exceeds these conditions and is thus generalizable to a wider context.

Lastly, let me briefly return to the debate between the scholars representing the "language affects cognition" account versus the ones defending the "language does not affect cognition" account. Given the marked congruency of the pattern of results of Experiments 1 and 2, there is clearly no evidence to falsify the theory of universal grammar (Chomsky, 1965). The language of the study materials—German or Italian—did not modulate the findings. Hence, my findings add support for sustaining Noam Chomsky's influential theory. Concomitantly, the findings provide no basis to support the Sapir-Whorf hypothesis of linguistic relativity (Sapir, 1921; Whorf, 1956). I want to clarify that this is obviously a highly tentative insight. It simply and solely contributes a tiny piece in the hope of continuing the debate, which of course remains an enigma.

3.3 Experiment 3

Experiment 3 was designed and conducted with two central aims in mind: First, one purpose was to provide a second replication for Experiments 1 and 2. Second, a further purpose of the present experiment was to extend previous findings by examining the role that relevance plays as a potentially important pragmatic moderator of the effects detected so far.

Even research that is characterized by excellence and the highest standards of scientific quality may produce unreliable empirical findings due to both systematic and random error (Open Science Collaboration, 2015). In many cases a single replication

study does not suffice. Maxwell et al. (2015) argued that an original study should be replicated more than once. A single replication study may be related to problems of statistical power. Imagine an original study finds an effect, which cannot be replicated in the first replication study. Of course, one could conclude that the finding from the original study was a false positive. However, one could just as well conclude that the lack of a finding in the replication study was a false negative, especially when considering statistical artefacts like regressive shrinkage (Fiedler & Prager, 2018). This implies that failures to replicate may not be failures at all, but rather are the consequence of low statistical power in a single replication study. This, in turn, invokes the need to conduct multiple replications. Hence, the likelihoods of a false positive original effect as well as a false negative replicated effect decrease. Shrout and Rodgers (2018) also plead for multiple replications to validate an effect, and draw attention to yet another advantage of this course of action: Multiple replications provide an array of effect sizes, which can be used to estimate the true effect by means of meta-analysis. As a consequence, the focus can shift to analyzing effect size distributions, rather than binary decisions based on *p*-values.

The extension of Experiment 3 in comparison to the previous two experiments refers to the incorporation of a concept that is paramount in order to advance our knowledge on the pragmatic underpinnings of rational argumentation with conditionals and counterarguments—relevance. The multi-layered meaning of this concept in psychological science implies different definitions and characterizations. Therefore, it is important to clearly work out how exactly relevance is construed in the present thesis, and demarcate it from other levels of interpretation. This helps to avoid conceptual ambiguities and equivocal nomenclature. Blanchette et al. (2014) classified three recognizable meanings of the relevance concept as it relates to the psychology of reasoning: (1) a formal definition, (2) a semantic definition, and (3) a goal-based

definition. According to the formal point of view, relevance is conceptualized as a utility function, meaning that relevance reflects the usefulness of information to the extent that it helps to solve a problem that demands the engagement of reasoning. Information that is not appropriate to determine whether a conclusion is valid or invalid is irrelevant. Consistent with this line of reasoning, Schaeken et al. (2007) have shown that people direct more attentional resources towards premises that are relevant compared with irrelevant ones. According to the semantic definition, relevance represents a relatedness function. As such, information ought to be semantically connected to a reasoning task in order for it to be deemed relevant. Thus, congruently contextualized information exerts higher relevance than does de-contextualized information (Evans, 1995, 1996, 2006). Thus, mental models with semantically related information are more probable to be fleshed out than are models that only entail irrelevant information (Johnson-Laird, 2006). According to the goal-based account, relevance is defined as a function of goal-achievement. Information is relevant as long as it helps reasoners progress towards their goals (Sperber & Wilson, 1986, 1995; Sperber et al., 1995). Girotto et al. (2001) demonstrated that reasoners primarily search for information that indicates which decision facilitates or hinders goal attainment, rather than information hinting at rule adherence. Similarly, research in social pragmatics has shown that information is interpreted as relevant when it is congruent with personal goals (e.g., Hilton et al., 2005). I decided to adopt the goal-based interpretative lens for the present study since it is closely tied to relevance theory (Sperber & Wilson, 1986, 1995) and a direct conceptual successor thereof.

While I have already described the central building blocks of relevance theory (Sperber & Wilson, 1986, 1995), its main assumptions, and its operating principles in the theory section of this thesis, I will now turn to the deduction of the rationale of Experiment 3. Concretely, when and how can relevance theory (Sperber & Wilson,

1986, 1995) be applied to rational argumentation with conditionals and counterarguments? And secondly, what predictions can be made for conditional reasoning with counterarguments based on relevance theory (Sperber & Wilson, 1986, 1995)? The major advantage of relevance theory (Sperber & Wilson, 1986, 1995) in comparison with other theories of reasoning is the fact that it transcends the boundaries and restrictions that other theories of reasoning impose on paradigm choice, design decisions, and procedural operationalizations. Other major theories of reasoning such as mental logic (Braine & O'Brien, 1991; Rips, 1994), mental models (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991), pragmatic schemas (Cheng & Holyoak, 1985; Cheng et al., 1986), and Darwinian algorithms (Cosmides, 1989) usually come along with specifically designated paradigmatic tests. While this approach is completely legitimate from an experimental point of view, it renders findings domain-specific and minimizes their scope and impact. Relevance theory (Sperber & Wilson, 1986, 1995), however, provides a comprehensive and substantive theoretical framework for explaining a large spectrum of pragmatic influences on reasoning. Thus, relevance theory as a theoretical framework is particularly suited to study conditional reasoning with counterarguments because the obtained results will very likely not merely apply to the exact conditions during experimentation, but will generalize across domains and hence allow for more general and robust conclusions regarding the interplay of conditional reasoning and counterarguments during rational argumentation.

Krzyżanowska et al. (2017) claim that a conditional is only assertable when the antecedent part is relevant for the consequent part of the conditional. If this boundary condition is not met, then conditionals render poor arguments (Oaksford & Hahn, 2007). That relevance relation between antecedent and consequent, however, is arguably modified by the relevance of the counterargument itself. That is, I conjecture

that the relevance of the counterargument interferes with the relevance link established between antecedent and consequent. More precisely, I assume that the relevance of the counterargument evokes a causal disruption of the relevance link between antecedent and consequent—hence reducing conclusion endorsement. By contrast, the relevance link between antecedent and consequent should remain intact when counterarguments are irrelevant—hence conclusion assertion would be unaffected. Consequently, the interplay of inference and mode is arguably only modulated by relevant counterarguments—not by irrelevant ones. Distinguishing between irrelevant and relevant counterarguments is also justified from a pragmatic perspective. Every speech act conveys a presumption of its own relevance (Sperber et al., 1995). It is this idea that Sperber and Wilson coined the principle of relevance (1986), and later referred to as the second (communicative) principle of relevance (1995). In fact, however, not every utterance is relevant. Indeed, a communicated piece of information may turn out not to be relevant at all. Likewise, a reasoner may decide to neglect an irrelevant counterargument for the actual inference process because the initial presumption—which has proven wrong—was refuted. Another reason to disregard irrelevant counterarguments before inferential integration starts is the tendency of reasoners to strive for minimal processing effort, at least under highly standardized conditions and during familiarized actions. Despite the fact that relevance theory constitutes an important framework for the study of pragmatics in philosophy, psychology, and linguistics, experimental work that tests its core assumptions and identifies important boundary conditions is scarce. Sperber et al. (1995) found in a series of four experiments with the Wason selection task (Wason, 1966) that expectations of relevance vary with rule and context, and thus does participants' reasoning performance. In line with this reasoning and transferred to the present research objective, it is highly predictable from a theoretical standpoint that the

relevance of a counterargument modulates the impact on conclusion endorsement exerted by the interplay of a conditional's inference type and its counterargument's linguistic mode during rational argumentation.

3.3.1 Hypothesis

In general, I tested whether the findings from the previous two experiments can be replicated a second time in Experiment 3. Additionally, Experiment 3 is an extension of the previous two experiments because it makes a novel and nuanced prediction regarding relevance as an important moderator of the findings. When counterarguments are absent, I expected that modus ponens inferences show higher endorsement ratings than modus tollens inferences. When counterarguments are present, I proposed two different sets of hypotheses depending on the relevance of the counterargument: For irrelevant counterarguments, I expected higher endorsement ratings for modus ponens inferences compared with modus tollens inferences. I did not expect an influence of mode for irrelevant counterarguments. Importantly, I hypothesized that irrelevant counterarguments do not instigate the previously observed interaction between inference and mode. However, for relevant counterarguments I anticipated the previously observed interaction between inference and mode to reoccur. Thus, I hypothesized that relevant counterarguments do instigate the interaction between inference and mode. Specifically, I hypothesized that given the presence of relevant counterarguments, for subjunctive mode endorsement ratings are higher for modus ponens inferences as opposed to modus tollens inferences, whereas this relationship reverses for indicative mode. Given the presence of relevant counterarguments, for indicative mode endorsement ratings are lower for modus ponens inferences as opposed to modus tollens inferences. In terms of response times, I grounded my predictions on the exploratory response time findings from Experiment 1 and the confirmatory response time findings of Experiment 2. When

counterarguments are absent, I hypothesized that response times are lower for modus ponens inferences than for modus tollens inferences. When counterarguments are present, I expected distinct effects of inference and relevance. Given the presence of counterarguments, overall modus ponens inferences should display lower response times compared with modus tollens inferences. I hypothesized that overall, response times are lower when irrelevant counterarguments are presented as opposed to showing relevant counterarguments. In line with the results from the previous two experiments, I assumed to observe null effects for the interaction between inference and mode in the response times, irrespective of relevance.

3.3.2 Method

This section gives a detailed description of the sample characteristics, study design, materials, and experimental procedures of Experiment 3.

Participants. An a-priori power analysis with G*Power (Faul et al., 2007, 2009) based on an unknown effect in a new design yielded a sufficient sample size of 14 participants under the assumptions of a type I error of $\alpha = .05$ and a power of $1 - \beta = .80$ in order to reliably detect a medium effect size of Cohen's $f = .25$ (Cohen, 1988; Ellis, 2010). However, as Experiment 3 was part of a multi-experiment session, 60 participants were scheduled in line with requirements of other experiments. Inclusion criterion for study eligibility was German as first language. Exclusion criterion was prior expertise in formal logic. One participant had to be excluded because German was not her first language. Two participants were excluded because after the practice trials they indicated that they did not memorize the instructions, which led the responsible research assistant to restart the experiment. One participant had to be excluded because the experimental program crashed. Hence, the final sample consisted of $N = 56$ participants, which were German university students of different majors who were between 20 and 60 years old ($M = 24.75$, $SD = 5.83$; 20 male). A sensitivity analysis

revealed that 56 participants are sufficient to reliably observe a minimum detectable effect (MDE) of Cohen's *f* = .12, given a type I error of α = .05 and assuming a power of 1 - β = .80. Participants were recruited by sending a circular email via the email server of the University of Tübingen, Germany. They were paid or received course credit and a chocolate bar for their participation.

Design. The experiment followed a 2 (inference: modus ponens vs. modus tollens) × 2 (counterargument: absent vs. present) × 2 (mode: subjunctive vs. indicative) × 2 (relevance: irrelevant vs. relevant) within-participants design, with the factors mode and relevance being nested in the level "present" of the factor counterargument. Thus, the experiment entailed 10 conditions. Dependent measures were endorsement rating and response time.

Materials. I created 60 scenarios in text form serving as reasoning tasks for Experiment 3, which referred to unequivocal, ordinary events from everyday life. Each scenario consisted of four stimuli: major premise, minor premise, counterargument, and conclusion. The major premise was always an if-then-clause, consisting of the antecedent (p) and the consequent (q). The minor premise displayed p for the "modus ponens" condition and ¬q for the "modus tollens" condition. The counterargument was either absent for the "absent" condition, or present for the "present" condition. In the latter, counterarguments were formulated in subjunctive mode for the "subjunctive" condition, or in indicative mode for the "indicative" condition. Likewise, counterarguments were either irrelevant for the "irrelevant" condition, or relevant for the "relevant" condition. The conclusion was phrased as question that asked for the response in the "modus ponens" condition and the "modus tollens" condition, respectively. All instructions (Appendix C1), stimuli (Appendix C2), and response formats (Appendix C3) of Experiment 3 are appended.

Procedure. After participants were welcomed and signed the informed consent form, they were seated in front of a computer screen in an individual lab room. The experiment was implemented and conducted using the program OpenSesame (Mathôt et al., 2012). A research assistant was present during the instructions and the practice trials to clarify potential questions before the main experiment started. First, participants were instructed about their task. They were informed that they will be presented reasoning tasks and that each reasoning task consists of several sentences. The first sentences are written in black font and include an if-then-clause, a fact, and possibly a counterargument. The last sentence is a question written in red font that asks for a conclusion. Participants were instructed to answer this question on a 7-point Likert scale using the numbers 1 to 7 written on green stickers, which were placed in one row on the number bar of the keyboard. Participants had to press the respective key to give their response. Key 1 represented “no, in no case”, while key 7 represented “yes, in any case”. For the counterbalanced version of the experiment, the scaling of the response format was reversed. Accordingly, key 1 represented “yes, in any case”, whereas key 7 represented “no, in no case”. Participants were instructed to give their responses based on how they would respond to these scenarios in everyday situations. Participants were informed that the sentences will appear sequentially and that they can proceed from one sentence to the next one by pressing the space bar. They were further informed that they can take a short break between tasks if needed and that they can continue by also pressing space bar. Participants were asked to read every reasoning task carefully since the sentences may also involve negations. If no questions remained, participants were asked to press space bar to start the practice phase of the experiment. The practice phase comprised five practice trials in order to familiarize the participants with their task. The practice trials were presented fully randomized. After the practice phase ended, participants were asked whether they

have any remaining questions. As soon as all questions (if any) were clarified, the research assistant left the room and the participants started the main phase of the experiment by pressing space bar. For each of the 10 experimental conditions, six scenarios were presented, resulting in 60 trials in total. The 60 trials were presented sequentially in fully randomized order. Each trial was presented once. Trials were given self-paced and without time restrictions. Participants proceeded from one trial to the next one by pressing space bar. Each trial comprised five stimuli that were presented sequentially in fixed order: a fixation point, a major premise, a minor premise, a counterargument (if applicable), and a conclusion. The conclusion was presented until the participants gave their response using the 7-point Likert scale on the keyboard. Stimuli appeared self-paced and without time restrictions. Participants proceeded from one stimulus to the next one by pressing space bar. Each stimulus was shown at the screen center of the monitor. The screen background was white. Participants received no feedback after giving a response. One half of the participants conducted the experiment with the default scaling of the response format, and the other half of the participants conducted the experiment with the counterbalanced scaling of the response format. Participants were alternately assigned to the default version or the counterbalanced version, respectively. After the main phase of the experiment ended, participants were asked to indicate their sex and age. Then, participants were thanked and asked to inform the research assistant. The research assistant double-checked and recorded whether or not participants' first language was German as well as whether or not participants had prior expertise in formal logic. Lastly, the participants were debriefed, compensated, thanked, and dismissed.

3.3.3 Results

Data sets of four participants had to be removed from the analysis due to reasons stated in the participants section of Experiment 3. Apart from that, all valid

data were included. The statistical software packages and the data pre-processing (i.e., structuring, cleaning, transforming, aggregating) were identical to the previous two experiments.

For both dependent measures, all statistical tests followed the same analytical strategy. First, I conducted a 2 (inference: modus ponens vs. modus tollens) repeated-measures ANOVA as univariate test in order to analyze the data in the conditions where counterarguments were absent. Then, I conducted a 2 (inference: modus ponens vs. modus tollens) × 2 (mode: subjunctive vs. indicative) × 2 (relevance: irrelevant vs. relevant) repeated-measures ANOVA as omnibus test in order to analyze the data in the conditions where counterarguments were present. In case of a significant three-way interaction in the omnibus test, I conducted two separate ANOVAs: a 2 (inference: modus ponens vs. modus tollens) × 2 (mode: subjunctive vs. indicative) repeated-measures ANOVA for the conditions with irrelevant counterarguments, and a 2 (inference: modus ponens vs. modus tollens) × 2 (mode: subjunctive vs. indicative) repeated-measures ANOVA for the conditions with relevant counterarguments. In case of significant interaction effects in the separate ANOVAs, I calculated paired samples *t*-tests as post-hoc comparisons in order to further analyze the exact nature of the differences. I applied Holm's method (Holm, 1979) to adjust *p*-values for multiple post-hoc comparisons. I report Bayes factors for all analyses. I report BF_{incl} for the repeated-measures ANOVAs, and BF_{10} for the paired samples *t*-tests. Besides, for the post-hoc comparisons of all significant two-way interaction effects of the separate ANOVAs, I provide a Bayesian sequential analysis representing how the evidence for H_1 or H_0, respectively, changes with increasing sample size.

Endorsement Ratings. The univariate test, a 2 (inference: modus ponens vs. modus tollens) repeated-measures ANOVA, revealed a significant main effect of inference, $F(1, 55) = 36.91$, $p < .001$, $\eta_p^2 = .40$, $BF_{incl} = 1.45e+5$. As hypothesized,

when counterarguments were absent endorsement ratings were higher for modus ponens inferences compared with modus tollens inferences. Figure 15 shows the mean endorsement ratings when counterarguments were absent for Experiment 3.

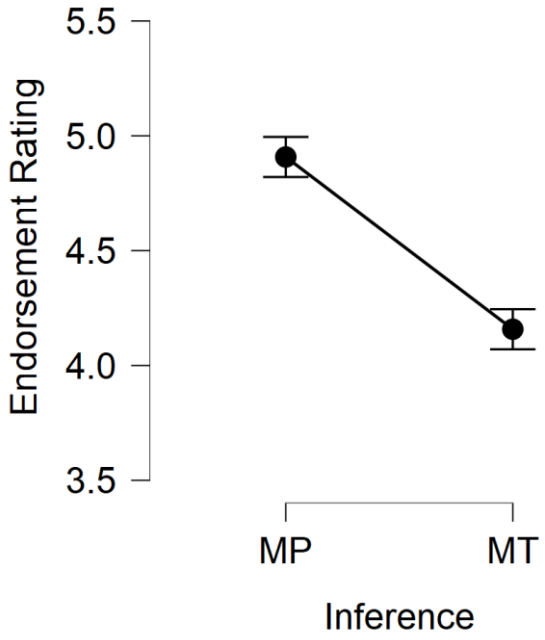

Figure 15. Mean endorsement ratings of Experiment 3 when counterarguments are absent as a function of inference. Inference: MP = Modus Ponens; MT = Modus Tollens. The scale ranges from 0 to 6. Error bars represent standard errors of the mean.

The omnibus test, a 2 (inference: modus ponens vs. modus tollens) × 2 (mode: subjunctive vs. indicative) × 2 (relevance: irrelevant vs. relevant) repeated-measures ANOVA, yielded a significant main effect of inference, $F(1, 55) = 15.39$, $p < .001$, $\eta_p^2 = .22$, $BF_{incl} = 2.35e+3$, a significant main effect of mode, $F(1, 55) = 39.42$, $p < .001$, $\eta_p^2 = .42$, $BF_{incl} = 1.73e+3$, and a significant main effect of relevance, $F(1, 55) = 346.52$, $p < .001$, $\eta_p^2 = .86$, $BF_{incl} = 8.84e+74$. The two-way interaction of inference and mode was significant, $F(1, 55) = 46.80$, $p < .001$, $\eta_p^2 = .46$, $BF_{incl} = 3.32e+2$. The two-way interaction of inference and relevance was also significant, $F(1, 55) = 56.03$, $p < .001$,

η_p^2 = .51, BF_{incl} = 2.64e+5. The two-way interaction of mode and relevance was significant as well, $F(1, 55) = 57.48$, $p < .001$, η_p^2 = .51, BF_{incl} = 1.47e+10. Most importantly, the three-way interaction of inference, mode, and relevance was significant, $F(1, 55) = 34.06$, $p < .001$, η_p^2 = .38, BF_{incl} = 2.98e+3. Figure 16 shows the mean endorsement ratings when counterarguments were present for Experiment 3.

The separate 2 (inference: modus ponens vs. modus tollens) × 2 (mode: subjunctive vs. indicative) repeated-measures ANOVA for irrelevant counterarguments yielded a significant main effect of inference, $F(1, 55) = 56.06$, $p < .001$, η_p^2 = .51, BF_{incl} = 9.20e+12. As predicted, endorsement ratings were higher for modus ponens inferences as opposed to modus tollens inferences when irrelevant counterarguments were presented. Interestingly, and contrary to my expectation of a null effect, the main effect of mode was significant, $F(1, 55) = 5.16$, $p = .027$, η_p^2 = .09, BF_{incl} = 1.51e+0. That is, endorsement ratings were lower for irrelevant counterarguments in subjunctive mode compared with irrelevant counterarguments in indicative mode. The interaction of inference and mode was not significant, $F(1, 55) = 0.23$, $p = .637$, η_p^2 = .00, BF_{incl} = 0.21e+0. As hypothesized, irrelevant counterarguments did not instigate an interaction effect of inference and mode.

The separate 2 (inference: modus ponens vs. modus tollens) × 2 (mode: subjunctive vs. indicative) repeated-measures ANOVA for relevant counterarguments did not indicate a significant main effect of inference, $F(1, 55) = 0.24$, $p = .629$, η_p^2 = .00, BF_{incl} = 0.16e+0. The main effect of mode was significant, $F(1, 55) = 81.96$, $p < .001$, η_p^2 = .60, BF_{incl} = 3.53e+11, suggesting that overall endorsement ratings were higher when relevant counterarguments were presented in subjunctive mode as opposed to indicative mode. However, as hypothesized this effect was qualified by inference, as evidenced by a significant interaction effect of inference and mode, F(1,

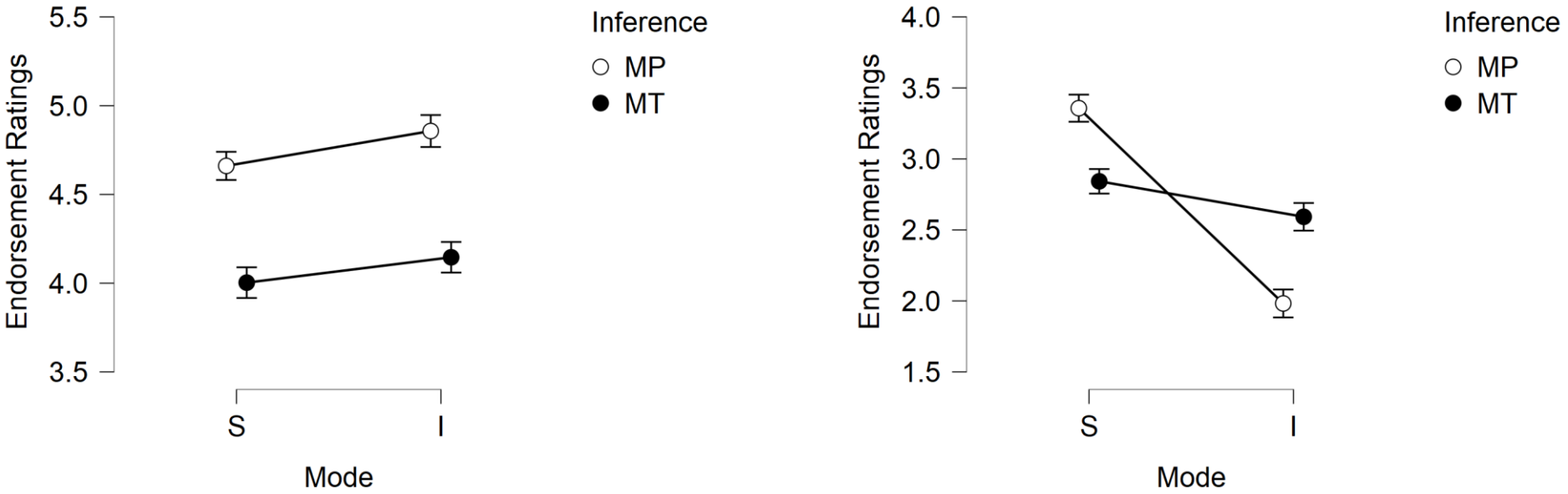

Figure 16. Mean endorsement ratings of Experiment 3 when counterarguments are present as a function of inference, mode, and relevance. Inference: MP = Modus Ponens; MT = Modus Tollens. Mode: S = Subjunctive; I = Indicative. Left panel: irrelevant counterarguments. Right panel: relevant counterarguments. The scale ranges from 0 to 6. Error bars represent standard errors of the mean.

55) = 63.54, $p < .001$, $\eta_p^2 = .54$, $BF_{incl} = 7.90e+6$. That is, relevant counterarguments did instigate an interaction effect of inference and mode. Paired samples *t*-tests show that endorsement ratings were significantly higher for modus ponens inferences with a relevant counterargument in subjunctive mode compared with modus ponens inferences with a relevant counterargument in indicative mode, $t(55) = 10.78$, $p < .001$, $d = 1.44$ (95% CI [1.06, 1.81]), $BF_{10} = 1.96e+12$. Endorsement ratings were also significantly higher for modus tollens inferences with a relevant counterargument in subjunctive mode compared with modus tollens inferences with a relevant counterargument in indicative mode, $t(55) = 2.52$, $p = .015$, $d = 0.34$ (95% CI [0.07, 0.61]), $BF_{10} = 2.61e+0$. As hypothesized, endorsement ratings were significantly higher for modus ponens inferences with a relevant counterargument in subjunctive mode compared with modus tollens inferences with a relevant counterargument in subjunctive mode, $t(55) = 4.39$, $p < .001$, $d = 0.59$ (95% CI [0.30, 0.87]), $BF_{10} = 4.07e+2$. As hypothesized, endorsement ratings were significantly lower for modus ponens inferences with a relevant counterargument in indicative mode compared with modus tollens inferences with a relevant counterargument in indicative mode, $t(55) = -4.92$, $p < .001$, $d = -0.66$ (95% CI [-0.94, -0.37]), $BF_{10} = 2.20e+3$. Figure 17 and Figure 18 show the respective Bayesian sequential analyses of the post-hoc comparisons.

Response Times. The univariate test, a 2 (inference: modus ponens vs. modus tollens) repeated-measures ANOVA, highlighted a significant main effect of inference, $F(1, 55) = 12.86$, $p < .001$, $\eta_p^2 = .19$, $BF_{incl} = 4.04e+1$. As hypothesized, when counterarguments were absent response times were lower for modus ponens inferences compared with modus tollens inferences. Figure 19 shows the mean response times when counterarguments were absent for Experiment 3.

The omnibus test, a 2 (inference: modus ponens vs. modus tollens) × 2 (mode: subjunctive vs. indicative) × 2 (relevance: irrelevant vs. relevant) repeated-measures

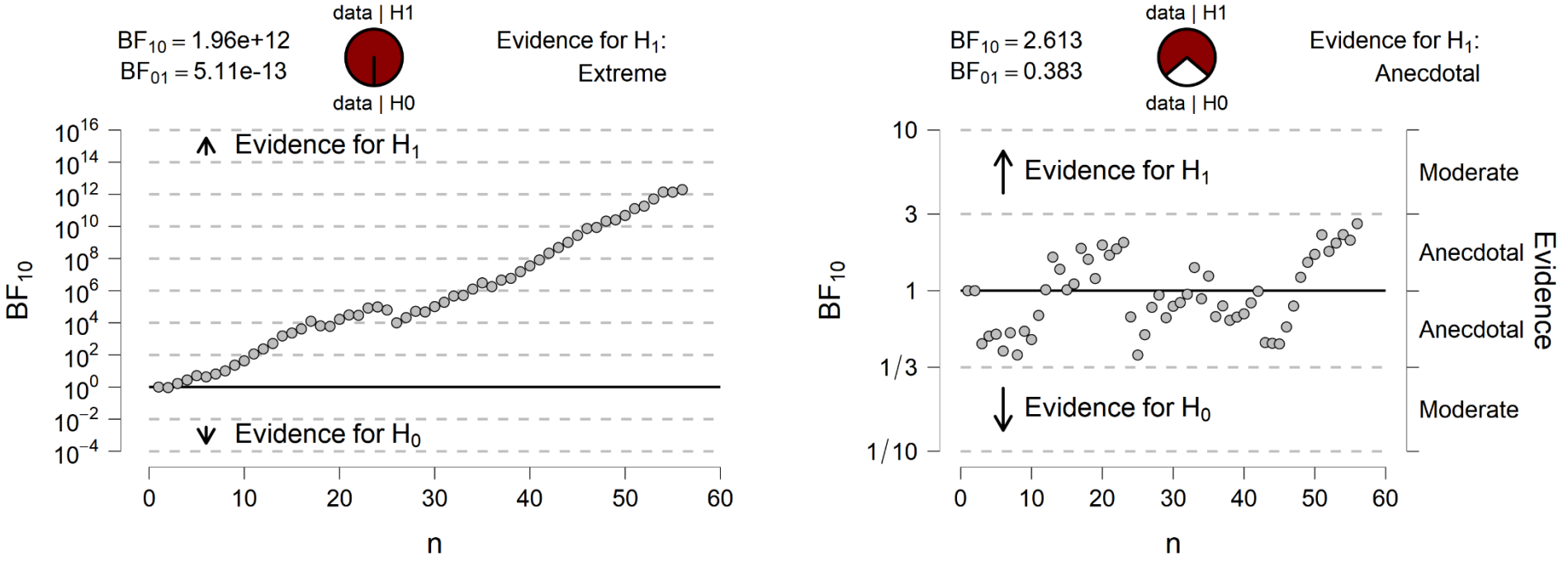

Figure 17. Bayesian sequential analysis of the endorsement ratings of Experiment 3 when counterarguments are present and relevant. Left panel: difference between MP inferences with subjunctive counterargument versus MP inferences with indicative counterargument. Right panel: difference between MT inferences with subjunctive counterargument versus MT inferences with indicative counterargument.

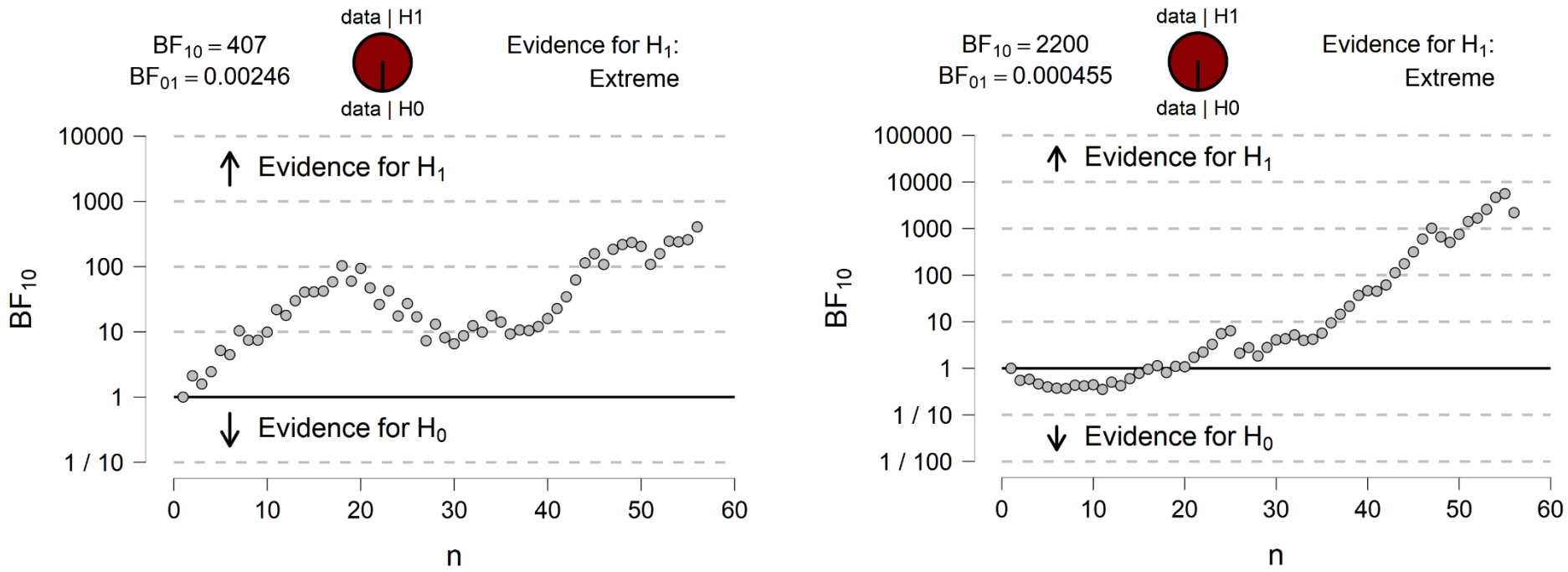

Figure 18. Bayesian sequential analysis of the endorsement ratings of Experiment 3 when counterarguments are present and relevant. Left panel: difference between MP inferences with subjunctive counterargument versus MT inferences with subjunctive counterargument. Right panel: difference between MP inferences with indicative counterargument versus MT inferences with indicative counterargument.

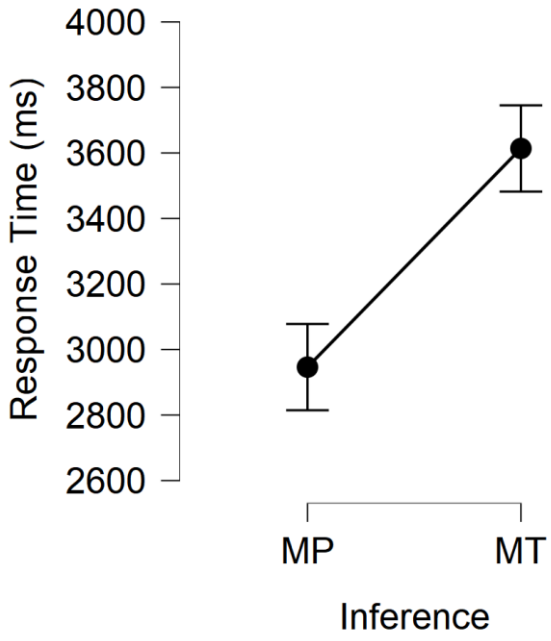

Figure 19. Mean response times of Experiment 3 when counterarguments are absent as a function of inference. Inference: MP = Modus Ponens; MT = Modus Tollens. Error bars represent standard errors of the mean.

ANOVA, displayed a significant main effect of inference, $F(1, 55) = 31.36$, $p < .001$, $\eta_p^2 = .36$, $BF_{incl} = 1.31e+9$. Expectedly, when counterarguments were present response times were overall lower for modus ponens inferences than modus tollens inferences. The main effect of mode was not significant, $F(1, 55) = 0.00$, $p = .975$, $\eta_p^2 = .00$, $BF_{incl} = 0.10e+0$. The main effect of relevance was significant, $F(1, 55) = 44.92$, $p < .001$, $\eta_p^2 = .45$, $BF_{incl} = 7.56e+7$. As predicted, when counterarguments were present response times were overall lower when the counterarguments were irrelevant rather than relevant. The two-way interaction of inference and mode was not significant, $F(1, 55) = 1.52$, $p = .223$, $\eta_p^2 = .03$, $BF_{incl} = 0.32e+0$. The two-way interaction of inference and relevance was also not significant, $F(1, 55) = 2.38$, $p = .129$, $\eta_p^2 = .04$, $BF_{incl} = 0.45e+0$. The two-way interaction of mode and relevance reached only marginal significance, $F(1, 55) = 3.94$, $p = .052$, $\eta_p^2 = .07$, $BF_{incl} = 0.58e+0$. As hypothesized, the three-way interaction of inference, mode, and relevance was not significant, $F(1, 55) = 2.79$, $p =$

.101, η_p^2 = .05, BF_{incl} = 0.47e+0, indicating that the interaction of inference and mode remained non-significant irrespective of relevance. Figure 20 shows the mean response times when counterarguments were present for Experiment 3.

3.3.4 Discussion

The aim of Experiment 3 was to conduct a second replication study that investigates the impact of inference type, counterarguments, and counterarguments' linguistic mode on conclusion endorsement in conditional reasoning. An extension of Experiment 3 was the examination of relevance as a crucial moderator of the observed effects. I will first summarize the findings and put a special focus on the modulations exerted by virtue of counterarguments' relevance. Moreover, I will provide an intriguing alternative explanation (in addition to the ones outlined earlier) for the consistent finding that modus ponens inferences are generally endorsed more frequently than modus tollens inferences, and that they require less time to be drawn. I will continue with an extensive discussion of the key concept motivating Experiment 3, namely relevance. To this end, the function of relevance as pragmatic context and its meaning for social rationality are illustrated. Finally, I invite the reader to an interesting excursion regarding the meaning of relevance for the cognition-emotion interface in order to reveal how relevance as a key concept in psychological science is also nomologically connected to more distal constructs.

The findings confirmed my hypothesis that modus ponens inferences are endorsed more frequently than modus tollens inferences. This was the case both for trials in which counterarguments were absent and for those in which irrelevant counterarguments were presented. This pattern of results was highly predictive given the findings of Experiments 1 and 2. However, I wish to propose an alternative account (next to the ones discussed before) to make sense of this finding. It has been argued that reasoners might display a verification bias when being confronted with conditionals

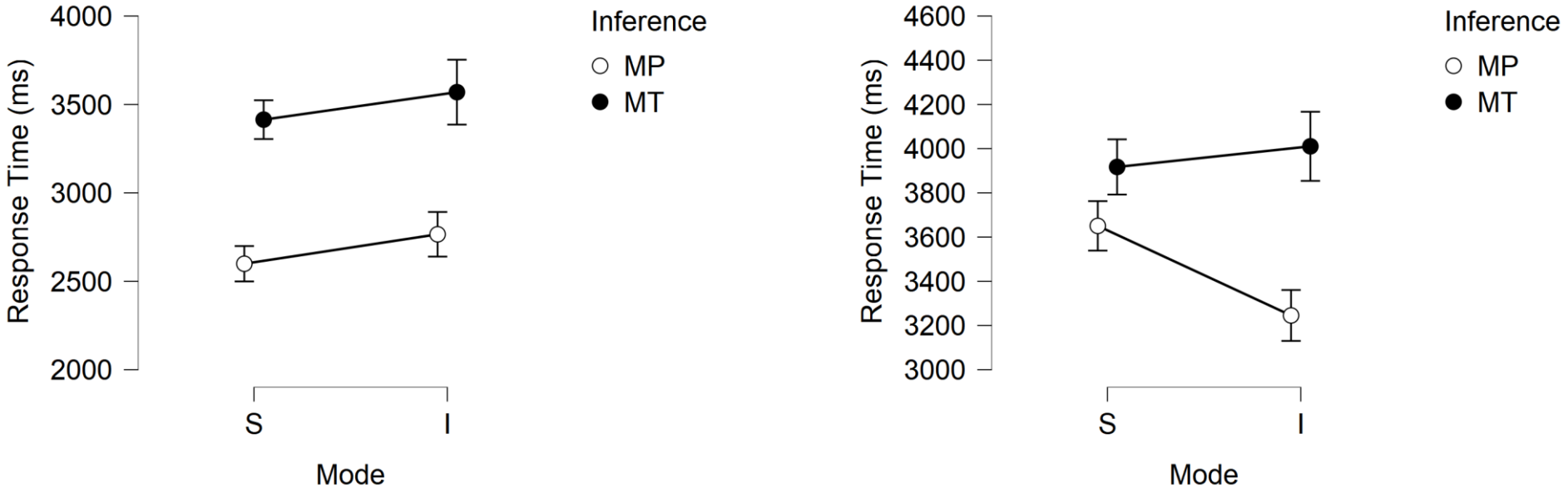

Figure 20. Mean response times of Experiment 3 when counterarguments are present as a function of inference, mode, and relevance. Inference: MP = Modus Ponens; MT = Modus Tollens. Mode: S = Subjunctive; I = Indicative. Left panel: irrelevant counterarguments. Right panel: relevant counterarguments. Error bars represent standard errors of the mean.

necessitating modus ponens versus modus tollens inferences (e.g., Lucas & Ball, 2005; Wason & Johnson-Laird, 1972). According to the notion of verification bias, people are prone to prove a rule true, rather than trying to prove it false. The inferential mechanism underlying modus ponens is a verification process, whereas the inferential mechanism underlying modus tollens is a falsification process. Hence, since people are subjected to verification bias, they more frequently draw the modus ponens inference by verifying the consequent as opposed to the modus tollens inference by falsifying the antecedent. But why is there a verification bias in conditional reasoning in the first place? Various theories propose various answers to this question. Mental model theory (Johnson-Laird, 1983, 2006; Johnson-Laird & Byrne, 1991, 2002) postulates that conditionals represent conjunctions of possibilities. Accordingly, the interpretation of a conditional corresponds to the construction of mental models, which represent possible states of affairs that are compatible with the conditional. Impossible states of affairs, in turn, are omitted (Johnson-Laird & Khemlani, 2013; Johnson-Laird et al., 2015; Khemlani et al., 2012, 2018; Skovgaard-Olsen, Collins, et al., 2019). The process of model construction, model inspection, and model variation requires working memory capacity. Since people have limited working memory capacity, they will represent as little as possible in their initial set of models to capture the meaning of the conditional. Therefore, people tend to only flesh out the affirmative rule inherent in modus ponens, whereas they do not explicitly represent a model for the negating rule immanent to modus tollens. Another way to make sense of verification bias in conditional reasoning is inspired by the heuristic-analytic theory (Evans, 1989, 1996). This theory posits that conditional reasoning involves two distinct processing stages—the heuristic stage and the analytic stage. The heuristic stage is preconscious and operates implicitly. During the heuristic stage, attention is selectively focused on task features that appear psychologically relevant. The analytic stage is conscious and

operates explicitly. During the analytic stage, higher-order cognitive effort is directed to these task features in order to draw the final conclusion. Evans (1998) argued that an attentional emphasis is applied towards affirmative aspects of the conditional during the heuristic stage, which has to do with the fact that negative terms in natural language do not alter the topic of an assertion. For instance, the statement “there is not an A” is still about the letter “A” rather than any other letter. This heuristic explains why there is a good level of logical performance for modus ponens inferences as opposed to modus tollens inferences—because more attentional focus is directed towards the affirmative elements rather than the negating aspects of a conditional. Then, the processing that takes place during the analytic stage only serves to rationalize the conclusion that has already been made on the basis of attentional capacity applied towards psychologically relevant aspects of the conditional during the heuristic stage (Evans, 1995; Lucas & Ball, 2005). A third approach to account for verification bias in conditional reasoning was forwarded by Oaksford and Chater (1994, 1995, 1996, 2003). The authors proposed a mathematical model called the optimal data selection account, which they derived from their general rational-analysis approach to human reasoning. Following optimal data selection, reasoners draw conclusions based on the information value of the premises in relation to their potential support for the conditional rule by calculating expected information gains. They mathematically demonstrated that maximal information gain is achieved by a data selection strategy that prevails to screen for affirmative rather than negating (but logically appropriate) elements of the premises. Thus, the optimal data selection account implies that the observed verification bias in conditional reasoning may in fact be a rational strategy in terms of probabilistic standards (Lucas & Ball, 2005). Irrespective of the functional details of the three theories, all three—mental model theory, heuristic-analytic theory, and optimal data selection theory—share an essential communality, namely the idea that negation

processes lead to increased processing effort, and this in turn makes modus tollens more difficult. The response time findings of Experiment 3 lend support for this idea: Modus tollens inferences generally consumed more time, both when counterarguments were absent and present.

Contrary to my expectation of a null effect, endorsement ratings were lower for irrelevant counterarguments in subjunctive mode compared with irrelevant counterarguments in indicative mode. This result might seem somewhat puzzling at first glance. Why should an indicative counterargument enable higher endorsements than a subjunctive counterargument, even if both counterarguments are irrelevant? From a pragmatic point of view, however, this pattern of results can be quite rational. Irrelevance can be clearly communicated in indicative mode. The recipient of the message can thus easily encode this counterargument and quickly extract the redundancy of its informational value with respect to the conditional. Consequently, irrelevant counterarguments in indicative mode are discarded early on in the inference process and thus do not affect the final conclusion as much. On the other side, irrelevant counterarguments in subjunctive mode arguably imply doubt as to whether or not the counterargument is really irrelevant for the conditional. Put differently, subjunctive mode may not decrease the belief in the irrelevance of the counterargument by changing its content, but it may launch the construction of new models representing what might be the case when the counterargument is false. Among these new models could be models that represent circumstances under which the otherwise valid conclusion to the conditional might be defeated. Nonetheless, I concede that this is a speculative post hoc interpretation of this specific finding, which would require further testing for its validation.

Most importantly, I found evidence that relevance functions as a central moderator for the interaction effect between inference and mode as observed in the

previous two experiments. For relevant counterarguments, the interaction remained intact: Endorsement ratings were higher for modus ponens compared with modus tollens in subjunctive mode; however, endorsement ratings were lower for modus ponens compared with modus tollens in indicative mode. Hence, I demonstrated this new finding for the third time. This increases my confidence in the existence of this effect. Therefore, I claim that I have revealed a novel psycholinguistic phenomenon in my doctoral work, a phenomenon that was utterly undiscovered before. Another objective of Experiment 3 was to identify boundary conditions of this effect. In line with my hypothesis, the interaction between inference and mode broke down when counterarguments were irrelevant. This highlights the fact that relevance constitutes a boundary condition for the inference-mode interaction in conditional reasoning with counterarguments. The response time findings provide auxiliary support for the important role of relevance. Overall, response times were higher when reasoners were confronted with relevant counterarguments rather than irrelevant ones. This suggests that participants more strongly integrated relevant counterarguments in the inference process in comparison to irrelevant counterarguments. This, in turn, presumably led to the pronounced interplay of inference and mode in such cases. Instead, irrelevant counterarguments were arguably neglected, or at least participants did not attribute as much informational value to them. Still pending are deliberations on the mediating principles driving the influence of relevance on conditional reasoning with counterarguments, which I will turn to now. Recent results (Skovgaard-Olsen, 2016b; Skovgaard-Olsen et al., 2016; Skovgaard-Olsen, Kellen, et al., 2017, 2019) suggest that relevance strongly moderates the evaluations of conditionals. These findings are based on ranking theory (Spohn, 2012, 2013). According to this extensive theoretical framework, a conditional expresses a reason relation between the antecedent and the consequent. This relationship is formalized probabilistically by the ΔP rule,

$$\Delta P = P(q|p) - P(q|\neg p). \quad (5)$$

In Equation 5, p denotes the antecedent and q denotes the consequent of the conditional. The term P(q|p) refers to the conditional probability of the consequent given the antecedent. The term P(q|¬p) denotes the consequent's conditional probability given the negated antecedent. Importantly, p is considered a reason for q, if p raises the probability of q, that is, if p is positively relevant for q. Since ΔP reflects the difference between P(q|p) and P(q|¬p), it must be positive for p to be a reason for q, and consequently for a conditional of the form "If p, then q" to be acceptable. The reason relation ΔP can result in three different outcomes,

$$P(q|p) - P(q|\neg p) > 0 \leftrightarrow \Delta P > 0, \quad (6)$$

$$P(q|p) - P(q|\neg p) = 0 \leftrightarrow \Delta P = 0, \quad (7)$$

$$P(q|p) - P(q|\neg p) < 0 \leftrightarrow \Delta P < 0, \quad (8)$$

where Equation 6 represents positive relevance, Equation 7 represents irrelevance, and Equation 8 represents negative relevance. For cases of positive relevance, Skovgaard-Olsen, Singmann, and Klauer (2017) showed that the conditional probabilities were a reliable predictor of the behavioral data mirroring the acceptability and the probability of the inferences. For cases of irrelevance and negative relevance, however, this relationship was disrupted because participants tended to view the conditionals as defective (Skovgaard-Olsen, Singmann, & Klauer, 2017). These findings suggest that a valid conditional indeed comprises a reason relation between antecedent and consequent, which explicitly manifests in the

relevance of the antecedent for the consequent. This conception of relevance is intrinsically probabilistic (e.g., see Oberauer et al., 2007; Over et al., 2007). Further empirical support for a probabilistic conceptualization of relevance comes from Krzyżanowska et al. (2017), who demonstrated that mere discourse coherence is not enough to render conditionals assertable; instead, probabilistic relevance is required in order to establish the causal link between antecedent and consequent. Other authors provided converging evidence for the notion that probabilistic relevance between cause and effect of a conditional constitutes a strong predictor of whether causal claims are classified as true or false (Berto & Özgün, 2021; Fitelson & Hitchcock, 2011; Pearl, 2000; Sikorski et al., 2019; Sprenger, 2018; Sprenger & Hartmann, 2019). The present findings from Experiment 3 contribute new insights into the working principles of probabilistic relevance in conditional reasoning. They indicate that the pragmatic context of conditionals—namely, the relevance of the counterargument—interferes with the relevance link between antecedent and consequent of the conditional. I suppose that it is precisely this mechanism that modulates the inference process, which eventually elicits distinct response patterns as a function of inference type and linguistic mode.

Undoubtedly, relevance constitutes a key player in the study of pragmatics (Austin, 1962; Grice, 1989). The present findings suggest that theories of reasoning must consider that relevance can be construed as pragmatic context. This pragmatic context can affect humans' conclusions and decisions during rational argumentation. A relevance-theoretic notion of rational argumentation acknowledges that relevance enriches the pragmatic context of an argument, whereas irrelevance can place constraints on it (Noh, 1996). Transferred to the present case, pragmatic enrichment by virtue of relevant counterarguments instigates a nuanced cognitive processing of the interplay between inference type of the conditional and linguistic mode of the

counterargument. Irrelevance, on the contrary, does not elicit this interaction because it lacks to provide a pragmatically enriching context for rational argumentation. Indeed, theorists plead for a much stronger emphasis on pragmatics to better understand rational argumentation (e.g., Bonnefon & Hilton, 2004; Evans & Over, 2004; Sperber et al., 1995; Thompson, 2000). Likewise, Sperber (2001) argues from an evolutionary perspective that the essential goal of all reasoning processes and argumentative speech acts is to manipulate the behavior and beliefs of other people. Following this rationale, the main function of rational argumentation is social in nature. Thompson et al. (2005) conducted a fascinating study demonstrating how closely rational argumentation is related to persuading and dissuading others. The authors of this study developed reasoning tasks with two different types of conditionals: persuasions (e.g., if the Kyoto accord is ratified, greenhouse gas emissions will be reduced) and dissuasions (e.g., if the Kyoto accord is ratified, there will be a downturn in the economy). Hence, the consequent was either offered as an incentive or disincentive for undertaking the antecedent. One group of participants was instructed to reason from the point of view of the writer of these statements; another group was instructed to reason from their own perspective. Interestingly, participants were more likely to adopt a deductive strategy when reasoning from the writer's point of view than their own point of view, even though no instructions to reason logically were included. This finding nicely demonstrates how most reasoning takes place in a pragmatic context. Reasoning seems to be most accurate and effective when searching for both evidence and counter-evidence for the persuasions and dissuasions of others during argumentation. Thompson et al. (2005) concluded:

> Thus, it seems unlikely that an adequate understanding of deductive processes will be achieved in the absence of an understanding of the role that deductive

> premises play in persuasive communications, and how these premises are interpreted and analyzed as part of an argument. (pp. 254–255)

Given that relevance plays an eminent part for the understanding of conditional reasoning in pragmatic contexts, I will now elaborate on the meaning of relevance for social rationality. First, a conditional is an extremely effective communicative device to express intentions and plans under conditions of uncertainty. The success of this communicative act is highly contingent on the solidity of the reasoned relationship among antecedent and consequent (i.e., the relevance of the antecedent for the consequent). This causal link, in turn, is affected by the relevance of potential counterarguments that function as epistemic modulators and regulate the pragmatic context of the conditional. Thus, relevance enables successful social coordination of joint organizational plans and shared goals between the speaker and the hearer (Hilton et al., 2005; Levinson, 1983, 2000). Relevance can therefore be defined as a form of social rationality that helps to enhance understanding between the sender and the recipient of a message. Relevance construed as social rationality is plausible from a phylogenetic perspective, too. According to this Darwinian account, evolutionary pressure has shaped modularized cognitive units in the human mind that served the specific adaptive function to monitor and regulate behaviors that align one's own goals with others' interests to engage in joint projects that yield collective benefits (Gärdenfors, 2003). For example, Fiddick et al. (2017) utilized repetition priming in conjunction with the Wason selection task to study whether reasoning relies on modularized units for specialized task (e.g., social contracts, precautions, deontic regulations). Their findings indeed supported the idea of specialized modules for specific reasoning abilities in different social domains (see also Cosmides, 1989; Gigerenzer & Hug, 1992). The results converge with early theoretical deliberations on

the functional architecture of the human mind, see for example Fodor's (1983) notion of a vertical faculty psychology. However, other authors raised concerns regarding the modularized organization and instead argue that social rationality is a product of general and domain-independent reasoning (e.g., Sperber, 2001; Vygotsky, 1962). It is clear that the debate on the functional organization of social rationality continues to be highly controversial. Still, what remains is the insight that relevance informs the construction of social rationality during argumentation.

Now, as a brief excursion, I wish to draw attention to the fact that relevance constitutes a powerful theoretical concept that does not solely influence language-based instances of reasoning. Relevance also affects the effects of emotion on thinking and reasoning. Despite the common belief that "[…] emotions are useless and bad for our peace of mind and our blood pressure" (Skinner, 1948, p. 92), as expressed by one of Burrhus Frederic Skinner's characters in the book *Walden Two*, it has since become widely accepted that emotion can both inhibit and facilitate sound inferences and rational thought. On the one hand, the induction of negative as well as positive mood can impede conditional reasoning (Oaksford et al., 1996), depressive mood reduces the frequency of correct responses during syllogistic reasoning (Channon & Baker, 1994; Radenhausen & Anker, 1988), a positive affective state as opposed to a neutral affective state reduces the amount of logical answers during reasoning with categorical syllogisms (Melton, 1995), thinking about emotional topics yields fewer normatively valid responses than does thinking about neutral topics during conditional reasoning problems (Blanchette, 2006; Blanchette & Richards, 2004), and decreased logicality was detected when reasoning about emotional content generally (Blanchette & Leese, 2011). On the other hand, experiments that utilized very intense stimuli and materials that were personally meaningful to participants have led to opposite results. Blanchette et al. (2007) showed that residents from London and Manchester in the

United Kingdom, who reported intensive emotional responses after the terror attacks in London in 2005, demonstrated higher syllogistic reasoning performance when tasks contained terrorism-related content compared against a Canadian participant sample, which was not involved in the terror attacks and which reported less intensive emotional responses. Another study showed that combat veterans reasoned better about emotional war-related topics than about neutral topics (Blanchette & Campbell, 2012). Investigations with victims of sexual abuse show increased performance on abuse-related contents (Blanchette & Caparos, 2013). Studies with patients suffering from various psychological disorders indicate an enhanced reasoning performance for content that relates to the specific disorder versus neutral content (Gangemi et al., 2014; Johnson-Laird et al., 2006). Blanchette et al. (2014) conducted a highly interesting study showing that the crucial moderator deciding whether emotion impedes or facilitates reasoning performance is relevance. The rationale underlying their study was the conjecture that relevance modulates the impact of emotion on reasoning in particular (Johnson-Laird & Oatley, 2000) and on cognitive functions in general (Blanchette & Richards, 2010). The authors used images and videos as stimuli, which were shown at once with or before the reasoning problems, respectively. The stimuli could be either neutral or emotional, and either irrelevant (i.e., semantically not related to the reasoning problem) or relevant (i.e., semantically related to the reasoning problem). Performance in reasoning was only negatively influenced by emotional contents that were irrelevant. However, relevant emotional content did not produce deleterious effects on reasoning. Blanchette et al. (2014) argued that this moderation by relevance is channeled via attentional mechanisms that can be explained through the semantic retrieval model (Forgues & Markovits, 2010; Markovits et al., 1998; Markovits & Quinn, 2002). This model implies that a semantic conditional automatically retrieves associated concepts from semantic memory via spreading activation. This, in

turn, puts an attentional focus on the reasoning task due to its semantic overlap with the retrieved concepts. Emotional content that is semantically unrelated, however, distracts the attentional focus away from the task and consequently impairs reasoning performance. Another idea refers to the powerful concept of affective meaning (Osgood, 1962, 1969; Osgood et al., 1957; Osgood et al., 1975). I have already shown, based on a rigorous Brunswikian sampling approach, that the dimensions of affective meaning—evaluation, potency, and activity—are learnt in distinct ways depending on the learning procedures employed. While stimulus pairing facilitates the acquisition of evaluation and activity, mere stimulus exposure affords the acquisition of potency and activity (Richter & Hütter, 2021). It would be an interesting endeavor to illuminate when and how relevance pervades these effects.

Taken together, I have accumulated ample evidence that allows for an interim conclusion. I have consistently demonstrated across three experiments that conditional reasoning is modulated by (1) the inference type of the conditional, (2) the presence of counterarguments, and (3) the linguistic mode of the counterarguments. I have further shown that (4) the obtained pattern of results replicates, and (5) is invariant of the language of the tested language-culture groups. Most importantly, I (6) identified relevance as an essential moderator and boundary condition of the findings, and thus (7) highlighted the function and meaning of relevance as pragmatic context for rational argumentation.

Following Sir Karl R. Popper (1935, 1963, 1972), I wish to emphasize that the strongest evidence for a theory comes from empirical corroborations of a priori hypotheses that are not readily deducible from rival theories. The predictions tested in my experiments, which predominantly could not be falsified, fall into this category. There are no previous data in the reasoning literature that point to the predicted relationships. Likewise, there is no rival theory that would generate similar predictions.

Instead, I based my predictions on careful a priori theorizing and a fine-grained conceptual analysis targeted at the identification of logical and plausible hypotheses that can be deductively derived from a theoretical integration of the core assumptions of relevance theory and the argumentative theory of reasoning.

3.4 Mixed Model Analysis

This chapter presents a mixed model analysis of the data I accumulated so far. The mixed model approach is based in large parts on work by Baayen et al. (2008). Mixed models, also referred to as multilevel models or hierarchical models, define a class of statistical models that account for the multi-layered data structure inherent to many experimental designs. These data structures include clustered data, repeated-measures data, longitudinal data, and clustered longitudinal data (Müller et al., 2013). Mixed models are powerful tools for these types of data (Baayen, 2008). Increasingly, they pave their way into experimental psychology (Hoffman & Rovine, 2007). Formally, a mixed model is a highly versatile extension of a multiple regression model (Maas & Hox, 2005). It includes both fixed effects parameters and random effects parameters. Fixed effects quantify the relationship of a predictor and a dependent variable for an entire population. Random effects quantify the relationship of specific clusters or subjects within a population by measuring the random variation in the dependent variable at different data levels (West et al., 2015). For instance, by setting the participant variable as random effects parameter, mixed models can control for idiosyncratic peculiarities of the participant sample. Consequently, more generalizable fixed effects estimates are computable (Singmann & Kellen, 2020).

A major advantage of mixed models over traditional statistical models is that they avoid an often neglected statistical pitfall: If a standard statistical method is applied to a data set with a multilevel structure, then the assumption of independent errors is violated (Nezlek, 2008). This leads to an underestimation of standard errors, which in

turn causes an overestimation of the test statistic. Obviously, this statistical pitfall increases the risk of type I errors (de Leeuw & Meijer, 2008; Holmes Finch et al., 2014; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999; Twisk, 2006). Further benefits of mixed models are the generalization of the results to other samples of participants (and stimuli), boosted measurement precision, an increase in statistical power, the coverage of otherwise unexplained variance in the data, the capacity to manage unbalanced data frames and incomplete data sets, the straightforward accommodation of metric, ordinal, and nominal predictors, and the general prevention of information loss (Judd et al., 2012, 2017). All these things considered, a mixed model analysis can help to gain confidence in the robustness of empirical findings whose data structure is hierarchical.

3.4.1 Hypothesis

The hypothesis of the mixed model analysis refers to the robustness of the empirical findings obtained in Experiment 1, Experiment 2, and Experiment 3. Aside from reproducibility and replicability, more recently robustness has been recognized as a further crucial concept to counteract the inflation of type I error rates. Reproducibility is defined as the test of the reliability of a previous finding with the same data and the same statistical analysis strategy. Replicability means testing the reliability of a prior finding with different data and the same statistical analysis strategy. Robustness means testing the reliability of a prior finding with the same data but a different statistical analysis strategy (Nosek et al., 2022). Robustness of empirical findings is important because statistical errors can add white noise and introduce bias, which can eventually prompt erroneous inferences (Shrout & Rodgers, 2018). For instance, Bakker and Wicherts (2011) examined 281 research articles. They found statistical errors in 18% of them. The vast majority of these errors made the findings more (apparently) significant than vice versa. Silberzahn et al. (2018) provided 29

analysis teams with the same set of data and were able to detect a formidable variance in the results (see also Botvinik-Nezer et al., 2020). It is clear that a fragile finding constitutes a risk factor for generalizability. Furthermore, a lack of robustness may suggest questionable practices like *p*-hacking or overfitting, whereby the credibility of such findings is tremendously reduced (Simonsohn et al., 2020; Steegen et al., 2016). Consequently, “[m]ore robust and general inferences must necessarily be supported by stronger statistical evidence” (Westfall et al., 2014, p. 2021). Hence, the central research question of the present chapter addresses the robustness of the findings from Experiments 1, 2, and 3. To this end, I reanalyzed all collected data using mixed models. I hypothesized that the findings from all experiments remain robust and are not meaningfully altered via this alternative statistical analysis strategy by and large.

3.4.2 Method

This section gives a concise overview of the statistical logic underlying the mixed model analysis. I employed linear mixed models because both outcome variables (endorsement ratings and response times) represented continuous data (for other data types, e.g., nominal data, generalized linear mixed models would be required). The predictors were modeled as fixed effects. Participants served as random effects variable and were modeled as random intercepts. Conceptually, trials (i.e., single observations) were modeled as a level 1 variable, which was nested within the level 2 variable participants. Thus, participants served as a contextual variable for grouping trials. This hierarchical data structure of the mixed model analysis allows to account for the variability across participants. Hereby, the random intercept variance estimate for participants is partialled out. The unaggregated trial-by-trial analysis prevents information loss. This way, mixed models provide improved analytical precision. Figure 21 depicts the hierarchical multilevel structure of the mixed model analysis (Field et al., 2012).

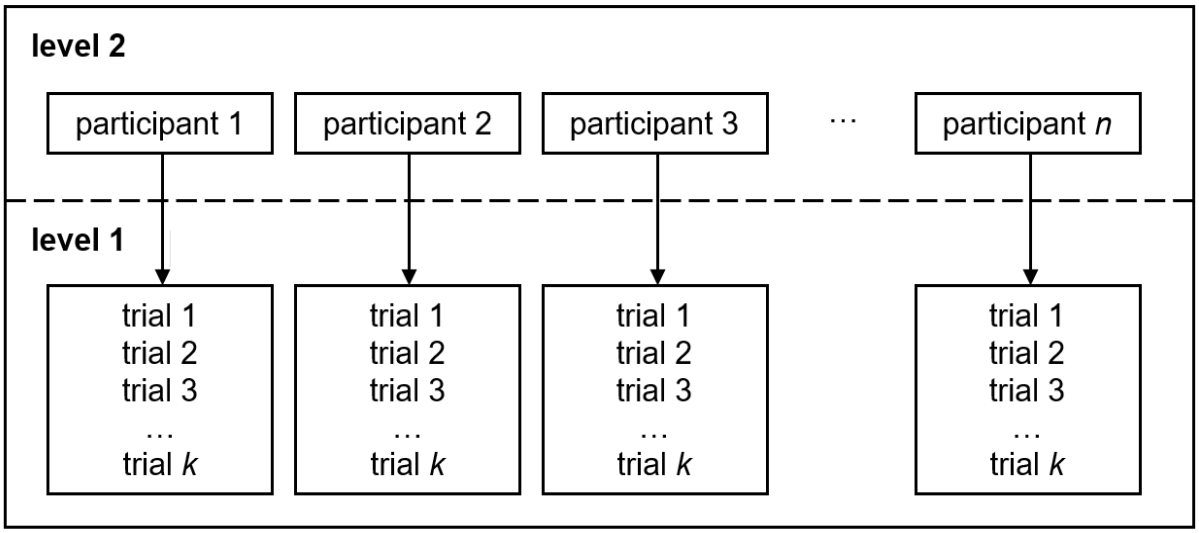

Figure 21. Hierarchical data structure of the mixed model analysis, adapted from Field et al. (2012, p. 857).

Note. n = number of participants. *k* = number of trials.

Formally, the mixed model analysis is a mathematical extension of a conventional regression analysis (Singmann & Kellen, 2020; Williams et al., 2021). It is mathematically expressed as follows:

$$y_{n,k} = a + u_n + b \cdot x_{n,k} + \varepsilon_{n,k}, \tag{9}$$

$$\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right), \tag{10}$$

$$u \sim N(0, \sigma_u^2). \tag{11}$$

In Equation 9, for the n^{th} participant and the k^{th} trial, y denotes the outcome, a denotes the intercept of the grand mean, u denotes the random effect that measures the variability among intercepts, b denotes the fixed effect, x denotes the predictor, and ε denotes the residual error component. Equation 10 shows that the residual error is assumed to follow a zero-centered Gaussian normal distribution with a variance

component σ_{ε}^2. Equation 11 shows that the random effect is assumed to follow a zero-centered Gaussian normal distribution with a variance component σ_u^2.

3.4.3 Results

Data were processed and analyzed using the statistical software R (Version, 4.2.0; R Core Team, 2022) and JASP (Version 0.16.1; JASP Team, 2022). I used the pre-processed (i.e., cleaned and transformed) data sets from Experiments 1, 2, and 3. Then, I restructured the data from wide format to long format. That is, endorsement ratings and response times were listed in an unaggregated fashion across single trials for each participant and each experimental condition. Accordingly, the data sets' rows represented single observations (one observation per trial); the data sets' columns represented participant, experimental conditions, and dependent measures.

For all three experiments and both dependent measures, I specified the following models: an omnibus model including all combinations of conditions, and separate models testing the 2 × 2 interactions. In addition, I specified a univariate model for Experiment 3 to cover the conditions in which counterarguments were absent.

For each model, I conducted the following statistical tests: First, I computed a linear mixed model with the experimental conditions as fixed effects and participant as random effect (i.e., random intercept). For each model, the number of observations *N* was calculated by multiplying the number of participants *n* with the number of trials *k*. Model terms were tested with likelihood ratio test (LRT; Wilks, 1938). Second, I assessed model fit using maximum likelihood estimation (MLE; Fisher, 1912). Third, the variance components (i.e., participant variance estimate and residual variance estimate) of the models were computed. Fourth, in case of significant interaction effects for the separate 2 × 2 models, I calculated pairwise contrasts as post-hoc comparisons,

based on the *z*-statistic and with adjusted *p*-values using Holm's method (1979), in order to further analyze the exact nature of the differences.

Mixed Model Analysis of Experiment 1: Endorsement Ratings. Table 2 shows the mixed model analysis for the endorsement ratings of Experiment 1. Table 3 shows the respective model fit indices. Table 4 shows the respective variance components.

Table 2. Mixed model analysis for the endorsement ratings of Experiment 1.

Model	χ^2	*df*	*p*
Omnibus			
Inference	9.87	1	.002**
Mode	145.01	2	<.001***
Inference × Mode	32.88	2	<.001***
None – Subjunctive			
Inference	38.69	1	<.001***
Mode	95.06	1	<.001***
Inference × Mode	0.54	1	.464
None – Indicative			
Inference	1.50	1	.221
Mode	121.29	1	<.001***
Inference × Mode	26.51	1	<.001***
Subjunctive – Indicative			
Inference	0.30	1	.583
Mode	5.62	1	.018*
Inference × Mode	19.55	1	<.001***

Note. The model terms were tested with likelihood ratio test (LRT).
*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Endorsement ratings were higher for modus ponens inferences without counterargument compared with modus ponens inferences with an indicative counterargument, $b = 2.66$, $z = 12.02$, $p < .001$. Endorsement ratings were higher for modus tollens inferences without counterargument compared with modus tollens

inferences with an indicative counterargument, b = 1.03, z = 4.64, p < .001. Endorsement ratings were higher for modus ponens inferences without counterarguments compared with tollens inferences without counterargument, b = 1.01, z = 4.56, p < .001. Endorsement ratings were lower for modus ponens inferences with an indicative counterargument compared with modus tollens inferences with an indicative counterargument, b = -0.63, z = -2.83, p = .005.

Endorsement ratings were higher for modus ponens inferences with a subjunctive counterargument compared with modus ponens inferences with an indicative counterargument, b = 1.09, z = 4.84, p < .001. Endorsement ratings did not differ between modus tollens inferences with a subjunctive counterargument and modus tollens inferences with an indicative counterargument, b = -0.33, z = -1.48, p = .139. Endorsement ratings were higher for modus ponens inferences with a subjunctive counterargument compared with modus tollens inferences with a subjunctive counterargument, b = 0.80, z = 3.55, p = .001. Endorsement ratings were lower for modus ponens inferences with an indicative counterargument compared with modus tollens inferences with an indicative counterargument, b = -0.63, z = -2.77, p = .011.

Table 3. Model fit indices for the endorsement ratings of Experiment 1.

Model	*N*	Deviance	Log-likelihood	AIC	BIC
Omnibus	720	2827.70	-1413.85	2843.70	2880.34
None – Subjunctive	480	1823.39	-911.69	1835.39	1860.43
None – Indicative	480	1909.15	-954.58	1921.15	1946.20
Subjunctive – Indicative	480	1929.57	-964.79	1941.57	1966.61

Note. The models were fitted using maximum likelihood estimation (MLE). *N* = number of observations.

Table 4. Variance components (as standard deviations) for the endorsement ratings of Experiment 1.

Model	SD_u	SD_ε
Omnibus	0.68	1.68
None – Subjunctive	0.69	1.56
None – Indicative	0.66	1.71
Subjunctive – Indicative	0.71	1.75

Note. SD_u = participant variance estimate; SD_ε = residual variance estimate.

Mixed Model Analysis of Experiment 1: Response Times. Table 5 shows the mixed model analysis for the response times of Experiment 1. Table 6 shows the respective model fit indices. Table 7 shows the respective variance components.

Table 5. Mixed model analysis for the response times of Experiment 1.

Model	χ^2	*df*	*p*
Omnibus			
Inference	17.93	1	<.001***
Mode	7.99	2	.018*
Inference × Mode	1.78	2	.411
None – Subjunctive			
Inference	14.23	1	<.001***
Mode	5.73	1	.017*
Inference × Mode	0.28	1	.597
None – Indicative			
Inference	10.89	1	<.001***
Mode	6.67	1	.010**
Inference × Mode	0.77	1	.381
Subjunctive – Indicative			
Inference	10.86	1	<.001***
Mode	0.16	1	.687
Inference × Mode	1.66	1	.198

Note. The model terms were tested with likelihood ratio test (LRT).
*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 6. Model fit indices for the response times of Experiment 1.

Model	*N*	Deviance	Log-likelihood	AIC	BIC
Omnibus	720	14530.26	-7265.13	14546.26	14582.89
None – Subjunctive	480	9775.67	-4887.84	9787.67	9812.71
None – Indicative	480	9537.68	-4768.84	9549.68	9574.73
Subjunctive – Indicative	480	9730.05	-4865.02	9742.05	9767.09

Note. The models were fitted using maximum likelihood estimation (MLE). *N* = number of observations.

Table 7. Variance components (as standard deviations) for the response times of Experiment 1.

Model	SD_u	SD_ε
Omnibus	2096.37	5692.51
None – Subjunctive	1864.40	6249.39
None – Indicative	1477.07	4875.09
Subjunctive – Indicative	2717.31	5876.68

Note. SD_u = participant variance estimate; SD_ε = residual variance estimate.

Mixed Model Analysis of Experiment 2: Endorsement Ratings. Table 8 shows the mixed model analysis for the endorsement ratings of Experiment 2. Table 9 shows the respective model fit indices. Table 10 shows the respective variance components.

Table 8. Mixed model analysis for the endorsement ratings of Experiment 2.

Model	χ^2	*df*	*p*
Omnibus			
Inference	2.02	1	.155
Mode	124.86	2	<.001***
Inference × Mode	13.36	2	.001***
None – Subjunctive			
Inference	12.61	1	<.001***
Mode	98.91	1	<.001***
Inference × Mode	0.01	1	.912

Model	χ^2	*df*	*p*
None – Indicative			
Inference	0.00	1	.960
Mode	98.70	1	<.001***
Inference × Mode	9.59	1	.002**
Subjunctive – Indicative			
Inference	0.02	1	.886
Mode	0.73	1	.392
Inference × Mode	9.19	1	.002**

Note. The model terms were tested with likelihood ratio test (LRT).
*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Endorsement ratings were higher for modus ponens inferences without counterargument compared with modus ponens inferences with an indicative counterargument, $b = 2.26$, $z = 9.62$, $p < .001$. Endorsement ratings were higher for modus tollens inferences without counterargument compared with modus tollens inferences with an indicative counterargument, $b = 1.23$, $z = 5.22$, $p < .001$. Endorsement ratings were marginally higher for modus ponens inferences without counterargument compared with tollens inferences without counterargument, $b = 0.53$, $z = 2.24$, $p = .051$. Endorsement ratings were marginally lower for modus ponens inferences with an indicative counterargument compared with modus tollens inferences with an indicative counterargument, $b = -0.51$, $z = -2.17$, $p = .051$.

Endorsement ratings were higher for modus ponens inferences with a subjunctive counterargument than for modus ponens inferences with an indicative counterargument, $b = 0.68$, $z = 2.76$, $p = .023$. Endorsement ratings did not differ between modus tollens inferences with a subjunctive counterargument and modus tollens inferences with an indicative counterargument, $b = -0.38$, $z = -1.55$, $p = .121$. Endorsement ratings were marginally higher for modus ponens inferences with a subjunctive counterargument than for modus tollens inferences with a subjunctive

counterargument, b = 0.56, z = 2.26, p = .072. Endorsement ratings were marginally lower for modus ponens inferences with an indicative counterargument than for modus tollens inferences with an indicative counterargument, b = -0.51, z = -2.05, p = .080.

Table 9. Model fit indices for the endorsement ratings of Experiment 2.

Model	*N*	Deviance	Log-likelihood	AIC	BIC
Omnibus	720	2924.57	-1462.29	2940.57	2977.21
None – Subjunctive	480	1879.16	-939.58	1891.16	1916.20
None – Indicative	480	1963.17	-981.58	1975.17	2000.21
Subjunctive – Indicative	480	1997.33	-998.66	2009.33	2034.37

Note. The models were fitted using maximum likelihood estimation (MLE). *N* = number of observations.

Table 10. Variance components (as standard deviations) for the endorsement ratings of Experiment 2.

Model	SD_u	SD_ε
Omnibus	0.53	1.81
None – Subjunctive	0.64	1.66
None – Indicative	0.63	1.82
Subjunctive – Indicative	0.32	1.92

Note. SD_u = participant variance estimate; SD_ε = residual variance estimate.

Mixed Model Analysis of Experiment 2: Response Times. Table 11 shows the mixed model analysis for the response times of Experiment 2. Table 12 shows the respective model fit indices. Table 13 shows the respective variance components.

Table 11. Mixed model analysis for the response times of Experiment 2.

Model	χ^2	*df*	*p*
Omnibus			
Inference	6.59	1	.010**
Mode	7.46	2	.024*
Inference × Mode	0.93	2	.630

Model	χ^2	*df*	*p*
None – Subjunctive			
Inference	7.43	1	.006**
Mode	7.75	1	.005**
Inference × Mode	1.23	1	.268
None – Indicative			
Inference	5.94	1	.015*
Mode	5.69	1	.017*
Inference × Mode	0.49	1	.485
Subjunctive – Indicative			
Inference	1.90	1	.168
Mode	0.01	1	.923
Inference × Mode	0.04	1	.852

Note. The model terms were tested with likelihood ratio test (LRT).
*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 12. Model fit indices for the response times of Experiment 2.

Model	*N*	Deviance	Log-likelihood	AIC	BIC
Omnibus	720	14396.29	-7198.15	14412.29	14448.93
None – Subjunctive	480	9424.61	-4712.30	9436.61	9461.65
None – Indicative	480	9613.21	-4806.60	9625.21	9650.25
Subjunctive – Indicative	480	9721.80	-4860.90	9733.80	9758.84

Note. The models were fitted using maximum likelihood estimation (MLE). *N* = number of observations.

Table 13. Variance components (as standard deviations) for the response times of Experiment 2.

Model	SD_u	SD_ε
Omnibus	2028.31	5179.40
None – Subjunctive	1796.67	4289.70
None – Indicative	1900.90	5245.70
Subjunctive – Indicative	2419.47	5848.63

Note. SD_u = participant variance estimate; SD_ε = residual variance estimate.

Mixed Model Analysis of Experiment 3: Endorsement Ratings. Table 14 shows the mixed model analysis for the endorsement ratings of Experiment 3. Table 15 shows the respective model fit indices. Table 16 shows the respective variance components.

Table 14. Mixed model analysis for the endorsement ratings of Experiment 3.

Model	χ^2	*df*	*p*
Univariate			
Inference	63.69	1	<.001***
Omnibus			
Inference	35.99	1	<.001***
Mode	36.66	1	<.001***
Relevance	891.51	1	<.001***
Inference × Mode	25.51	1	<.001***
Inference × Relevance	47.46	1	<.001***
Mode × Relevance	84.79	1	<.001***
Inference × Mode × Relevance	30.84	1	<.001***
Irrelevant			
Inference	102.72	1	<.001***
Mode	6.55	1	.010**
Inference × Mode	0.16	1	.686
Relevant			
Inference	0.55	1	.549
Mode	100.38	1	<.001***
Inference × Mode	49.09	1	<.001***

Note. The model terms were tested with likelihood ratio test (LRT).
*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Endorsement ratings were higher for modus ponens inferences with a relevant counterargument in subjunctive mode compared with modus ponens inferences with a relevant counterargument in indicative mode, $b = 1.38$, $z = 12.23$, $p < .001$. Endorsement ratings were higher for modus tollens inferences with a relevant

counterargument in subjunctive compared with modus tollens inferences with a relevant counterargument in indicative mode, $b = 0.25$, $z = 2.22$, $p = .026$. Endorsement ratings were higher for modus ponens inferences with a relevant counterargument in subjunctive mode compared with modus tollens inferences with a relevant counterargument in subjunctive mode, $b = 0.52$, $z = 4.58$, $p < .001$. Endorsement ratings were lower for modus ponens inferences with a relevant counterargument in indicative mode compared with modus tollens inferences with a relevant counterargument in indicative mode, $b = -0.61$, $z = -5.43$, $p < .001$.

Table 15. Model fit indices for the endorsement ratings of Experiment 3.

Model	*N*	Deviance	Log-likelihood	AIC	BIC
Univariate	672	2221.73	-1110.87	2229.73	2247.77
Omnibus	2688	9451.89	-4725.94	9471.89	9530.85
Irrelevant	1344	4453.58	-2226.79	4465.58	4496.80
Relevant	1344	4919.27	-2459.63	4931.27	4962.49

Note. The models were fitted using maximum likelihood estimation (MLE). *N* = number of observations.

Table 16. Variance components (as standard deviations) for the endorsement ratings of Experiment 3.

Model	SD_u	SD_ε
Univariate	0.64	1.19
Omnibus	0.58	1.37
Irrelevant	0.68	1.21
Relevant	0.61	1.46

Note. SD_u = participant variance estimate; SD_ε = residual variance estimate.

Mixed Model Analysis of Experiment 3: Response Times. Table 17 shows the mixed model analysis for the response times of Experiment 3. Table 18 shows the respective model fit indices. Table 19 shows the respective variance components.

Table 17. Mixed model analysis for the response times of Experiment 3.

Model	χ^2	*df*	*p*
Univariate			
Inference	15.68	1	<.001***
Omnibus			
Inference	61.38	1	<.001***
Mode	0.00	1	.976
Relevance	53.62	1	<.001***
Inference × Mode	2.10	1	.147
Inference × Relevance	3.04	1	.081
Mode × Relevance	3.55	1	.060
Inference × Mode × Relevance	2.30	1	.130
Irrelevant			
Inference	46.85	1	<.001***
Mode	1.89	1	.170
Inference × Mode	0.00	1	.962
Relevant			
Inference	18.30	1	<.001***
Mode	1.68	1	.195
Inference × Mode	4.30	1	.038*

Note. The model terms were tested with likelihood ratio test (LRT).
*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Response times were marginally higher for modus ponens inferences with a relevant counterargument in subjunctive mode compared with modus ponens inferences with a relevant counterargument in indicative mode, $b = 405.23$, $z = 2.39$, $p = .051$. Response times did not differ between modus tollens inferences with a relevant counterargument in subjunctive and modus tollens inferences with a relevant counterargument in indicative mode, $b = -93.41$, $z = -0.55$, $p = .583$. Response times did not differ between modus ponens inferences with a relevant counterargument in subjunctive mode and modus tollens inferences with a relevant counterargument in subjunctive mode, $b = -266.61$, $z = -1.57$, $p = .233$. Response times were lower for modus ponens inferences with a relevant counterargument in indicative mode

compared with modus tollens inferences with a relevant counterargument in indicative mode, b = -765.25, z = -4.50, p < .001.

Table 18. Model fit indices for the response times of Experiment 3.

Model	*N*	Deviance	Log-likelihood	AIC	BIC
Univariate	672	12296.12	-6148.06	12304.12	12322.17
Omnibus	2688	49096.53	-24548.26	49116.53	49175.49
Irrelevant	1344	24539.97	-12269.99	24551.97	24583.19
Relevant	1344	24619.89	-12309.95	24631.89	24663.11

Note. The models were fitted using maximum likelihood estimation (MLE). *N* = number of observations.

Table 19. Variance components (as standard deviations) for the response times of Experiment 3.

Model	SD_u	SD_ε
Univariate	904.98	2171.18
Omnibus	1086.22	2179.71
Irrelevant	1009.81	2147.70
Relevant	1173.92	2202.63

Note. SD_u = participant variance estimate; SD_ε = residual variance estimate.

3.4.4 Discussion

The objective of this chapter was to provide a mixed model analysis of the accumulated evidence in order to test whether the findings remain robust when accounting for random variation in responses between participants. To this end, the participant variable was modeled as random intercept. Confirming my hypothesis, the results obtained via the mixed model analysis were virtually congruent with the findings of the conventional statistical tests that I conducted prior. I will focus the discussion of the mixed model results on the central findings of this thesis in order to underscore their robustness. Regarding the endorsement ratings, I discuss the two-way interaction between inference (modus ponens vs. modus tollens) and mode (subjunctive vs.

indicative), as well as its moderation by relevance. I will also highlight the robustness of the response times. Finally, I will discuss confinements and potential extensions of the mixed model analysis.

Importantly, the two-way interaction effect between inference (modus ponens vs. modus tollens) and mode (subjunctive vs. indicative) was significant again. Endorsement ratings were higher for modus ponens inferences with counterarguments in subjunctive mode compared with modus tollens inferences with counterarguments in subjunctive mode. Conversely, endorsement ratings were lower for modus ponens inferences with counterarguments in indicative mode compared with modus tollens inferences with counterarguments in indicative mode. This pattern re-emerged across all experiments. Crucially, the mixed model analysis of Experiment 3 repeatedly demonstrated that this interaction effect was moderated by relevance. While relevant counterarguments instigated the observed outcome, irrelevant counterarguments did not (i.e., the interaction effect collapsed). Also in line with the findings from the previous analyses, the two-way interaction between inference (modus ponens vs. modus tollens) and mode (subjunctive vs. indicative) and its modulation by relevance was not mirrored in the response times. The response time findings of the mixed model analysis once more revealed that, overall, modus ponens inferences required less time than modus tollens inferences. Furthermore, counterarguments increased response time, and irrelevant counterarguments required less time than relevant counterarguments to be processed. Taken together, the mixed model analysis demonstrated the reliability of the findings by verifying their robustness across different statistical procedures. Hence, the pattern of results is not a statistical artefact that is due to a violation of the assumption of independence of the residual errors and thus an underestimation of the standard errors. This renders it highly unlikely for the results to be false positives. Similarly, the mixed model analysis suggests that the findings are most likely not

caused by random variation of responses between participants due to idiosyncratic features of the sample. Therefore, the accumulated evidence is highly predictive of the response patterns of other samples of participants.

I included participants as random effects in the mixed model analysis reported here. This is in line with recent calls to emphasize idiosyncratic features in psychological research (Hamaker, 2012; Molenaar, 2004). The main argument is that the consideration of participant variance can help us to gain a richer understanding of psychological processes after all (Williams et al., 2021). I confined the mixed models to model the participant variable as random intercept. Of course, as mixed models are highly flexible statistical tools, participants could also be modeled as random slopes, or both as random intercepts and random slopes. I decided to model participants as random intercepts because it is widely accepted and common practice to include random intercepts. The inclusion of random slopes, however, oftentimes produces complications at the cost of model parsimony. For example, it is not seldom that the simultaneous modeling of random intercepts and random slopes leads to problems of non-convergence and overfitting (Westfall et al., 2014). Since I conducted multiple mixed models, which automatically increases the risk of such issues to occur, I decided to stick to the most robust and most commonly practiced approach to model between-participant variation. Another confinement of the present models is that I only modeled participants as random effects—not stimuli. Of course, stimuli could be treated as random effects, too. However, this may severely raise model complexity and produce singular model fit, i.e., the specified random effects parameters cannot be estimated accurately from the available data. It is advised to carefully reduce the random effects structure in such cases. A helpful heuristic is that one should generally avoid fitting overly complex models; however, the variance-covariance matrices should be estimated with sufficiently high precision (Matuschek et al., 2017). A careful and

reasonable model selection must take into account the tradeoff between predictive precision and overfitting (Bates et al., 2015). Consequently, I decided to model random effects for participants but not for stimuli. Moreover, this approach allowed for a consistent reporting of the modeling results. Given that the inclusion of random effects for stimuli would have led to overfitting for some models (while for others not), a comparison between the results of these models would have been hardly feasible. Instead, keeping the models more parsimonious allowed for consistent reporting and robust comparisons between the effects of the different models. Moreover, the construction principle of the stimulus materials followed the rationale of creating stimuli that resemble common scenarios from everyday life. I refrained from using peculiar thematic subdomains of stimuli, which would have been much more likely to introduce random variation in the items. Instead, I used more generic and commonly familiar themes. This suggests that the findings do most probably not depend on specific peculiarities of the sample of stimuli. Rather, it is plausible to predict that a similar pattern of results will emerge in other stimulus samples.

Lastly, I wish to mention that I used linear mixed models in the present analysis. Another powerful, alternative model class is generalized additive mixed modeling (GAMM; Baayen et al., 2017). GAMMs deal with the temporal autocorrelational structure of experimental data. This autocorrelational structure represents the attentional oscillations through the course of an experiment that are due to learning and fatigue. I consider it an interesting task for future research to take an even closer look at the psycholinguistic data reported in this thesis by means of GAMMs.

3.5 Meta-Analysis

This chapter features a meta-analysis of the cumulative evidence reported in this thesis. In 1976, Glass introduced the notion of meta-analysis as “[…] the statistical analysis of a large collection of analysis results from individual studies for the purpose

of integrating the findings" (p. 3). Shortly after this publication, other authors presented similar statistical approaches for research synthesis and the generalization of research findings (Rosenthal & Rubin, 1978; Schmidt & Hunter, 1977). Since then, meta-analysis has become a widespread methodology in the cognitive, behavioral, and social sciences (e.g., Borenstein et al. 2021; Cooper et al., 2009; Durlak & Lipsey, 1991; Glass et al., 1981; Hedges, 1992; Lipsey & Wilson, 1993, 2001; Wachter & Straf, 1990). A meta-analysis can be broadly defined as a statistical tool that provides a systematic quantitative approach to evidence synthesis. It estimates a combined effect size (and variance) aggregated from a sample of individual studies that address the same research question or test the same hypothesis, respectively (Field & Gillett, 2010). Thus, meta-analysis affords generalized inferences and facilitates a conclusive interpretation of empirical findings (Cooper et al., 2009).

A major advantage of meta-analysis is that it concentrates on actual effect sizes and therefore offers a strong alternative to the reliance on statistical significance. The mere focus on statistical significance is flawed. It often tells us little or nothing about the practically meaningful effect of an experimental manipulation, an educational intervention, or a clinical treatment. Instead, a meta-analysis mirrors the cumulative character of the genesis of scientific knowledge. It constitutes a hopeful way forward because it provides a sound basis on which to corroborate, refute, or modify theories. In this spirit, it concurs with Popper's (1935, 1945, 1963, 1976, 1994) epistemological philosophy of critical rationalism. Another stellar scholar, Paul E. Meehl, also argued in a seminal and fascinating article from 1978 that "[…] significant differences are little more than complex, causally uninterpretable outcomes of statistical power functions" (p. 806). Meta-analyses exceed these limitations of significance testing.

Further benefits of meta-analyses that support a good way of doing science are

as follows: Meta-analyses surpass single studies with respect to validity generalizations because they enable a test of whether an empirical finding can be obtained across various study settings, times, participant populations, and researchers (Cooper et al., 2009). Meta-analyses also increase measurement precision and reliability, for instance by providing confidence intervals around the effect size estimates. Meta-analyses help to identify outliers and influential cases among the included primary studies (Viechtbauer & Cheung, 2010), and consequently help to hedge against type I errors as well as type II errors (Goh et al., 2016). Lastly, meta-analysis as a systematic quantitative approach for evidence synthesis offers a more objective method than narrative reviews (which are also valuable tools for research integration, but are less standardized and therefore inevitably more difficult to compare with one another).

3.5.1 Hypothesis

Across three experiments and an additional mixed-model analysis, I obtained consistent evidence for the primary hypothesis of this thesis: The inference type of a conditional and the linguistic mode of a counterargument interactively predict conclusion endorsement. Specifically, conclusion endorsement is higher for modus ponens inferences as opposed to modus tollens inferences when subjunctive counterarguments are presented, whereas it is lower when indicative counterarguments are presented. The respective response time findings revealed that this pattern was associated with lower response times for modus ponens inferences than for modus tollens inferences, irrespective of the linguistic mode of the counterargument. The research aim of the present chapter addresses the robustness of these findings. To this end, I conducted a series of internal restricted-maximum-likelihood random-effects meta-analyses to assess the robustness of these findings on endorsement ratings and response times across the experiments. I hypothesized that

the aforementioned effects are homogenous across all three experiments. I further hypothesized that the aforementioned effects are validated by meta-analytically combined effect size estimates that significantly differ from zero in the predicted directions.

3.5.2 Method

This section provides a concise overview of the statistical logic underlying the meta-analysis. Principally, when performing a meta-analysis the investigator must make an informed choice on whether to base the meta-analysis on a fixed effects model or a random effects model (Hedges, 1992). The model selection is critically important for the scope of the inferences one intends to draw as well as the context to which one wants to generalize (Borenstein et al., 2010). The fixed effects model assumes that the observed effects are sampled from one distribution with one true effect. Accordingly, the observed effect size estimates vary only due to within-study error. A fixed effects model makes sense when there are good reasons to assume that the studies are functionally identical and the goal is to draw inferences that refer to the specific population only. In contrast, the random effects model assumes that the observed effects are sampled from several distributions with several true effects. Accordingly, the observed effect size estimates vary due to within-study error and between-study error (Borenstein et al., 2021; Grasman, 2017). The goal of the random effects model is to generalize to a range of populations.

In most cases, there is no reason to believe that the true effect size is exactly the same in all studies included in a meta-analysis because study settings and participant pools can vary between studies. This renders the random effects model the appropriate choice. Furthermore, the random effects model is more likely to fit the actual sampling distribution, does not impose a restriction of a common effect size, yields the identical results as the fixed effects model in the absence of heterogeneity,

and allows the conclusions to be generalized to a wider array of situations (Borenstein et al., 2010). Since I implemented different experimental designs across the three experiments, drew my samples from different participant pools, and aimed to draw strong and generalizable conclusions that extrapolate beyond different scenarios, I utilized a random effects model to conduct the meta-analysis (see Figure 22).

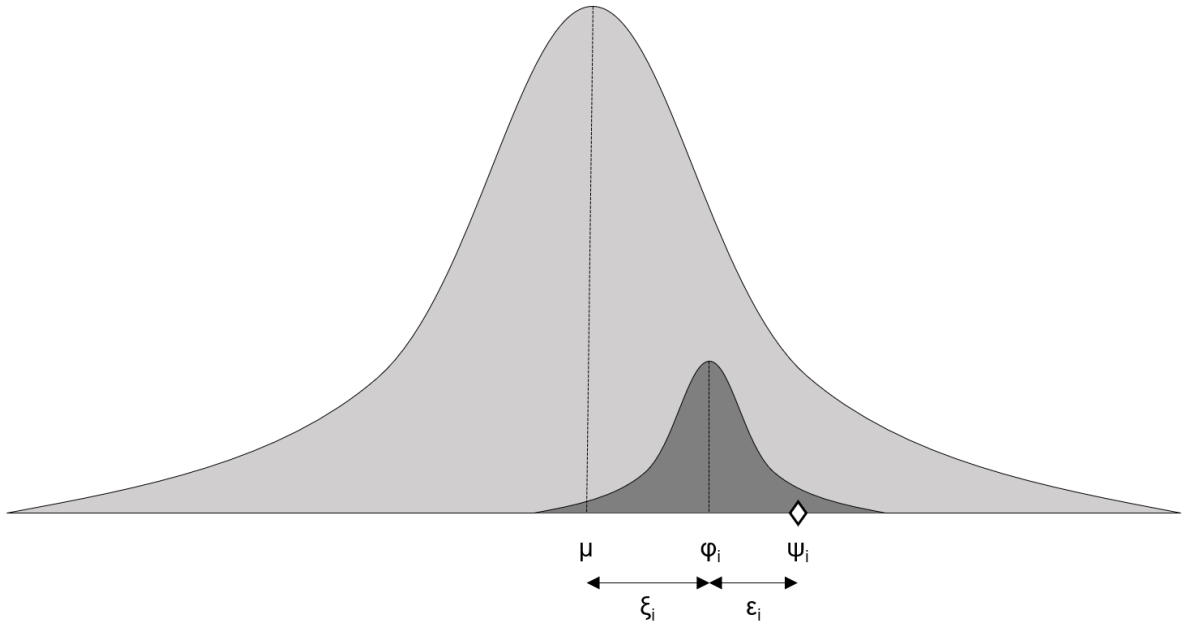

Figure 22. Distributional schematic of the random effects model for the meta-analysis, adapted from Borenstein et al. (2010, p. 100).

The random effects model is mathematically expressed as follows:

$$\psi_i = \varphi_i + \varepsilon_i, \quad (12)$$

$$\varphi_i = \mu + \xi_i, \quad (13)$$

$$\psi_i = \mu + \xi_i + \varepsilon_i. \quad (14)$$

For study i, ψ denotes the observed effect and φ denotes the true effect. The mean of all true effects is reflected by μ. The within-study error is represented by ε. The between-study error is represented by ξ. Equation 12 shows that the observed effect is an additive function of the study's true effect and the within-study error. Equation 13 shows that the study's true effect is an additive function of the mean of all true effects and the between-study error. Consequently, Equation 14 shows that the observed effect is an additive function of the mean of all true effects, the between-study error, and the within-study error.

3.5.3 Results

Data were processed using the statistical software R (Version 4.2.0; R Core Team, 2022) and analyzed with the R package metafor (Version 3.4-0; Viechtbauer, 2010a, 2010b). I used the pre-processed (i.e., cleaned, transformed, and aggregated) data sets from Experiments 1, 2, and 3. Then, I computed the means and standard deviations for each experimental condition of each experiment and for both dependent measures. I imported these means and standard deviations together with the respective sample sizes into data frames to calculate the corresponding effect sizes and variances. A random effects model was specified for each internal meta-analysis using the restricted maximum likelihood method (REML; Cooper & Thompson, 1977; Harville, 1977; Patterson & Thompson, 1971; Verbyla, 1990).

I conducted four internal restricted-maximum-likelihood random-effects meta-analyses: Meta-analysis 1 assessed the combined effect size estimate for the difference in endorsement ratings between modus ponens inferences with subjunctive counterarguments versus modus tollens inferences with subjunctive counterarguments. Meta-analysis 2 assessed the combined effect size estimate for the difference in endorsement ratings between modus ponens inferences with indicative counterarguments versus modus tollens inferences with indicative counterarguments.

Meta-analysis 3 assessed the combined effect size estimate for the difference in response times between modus ponens inferences with subjunctive counterarguments versus modus tollens inferences with subjunctive counterarguments. Meta-analysis 4 assessed the combined effect size estimate for the difference in response times between modus ponens inferences with indicative counterarguments versus modus tollens inferences with indicative counterarguments.

For each meta-analysis, the following statistical tests were computed: First, model fit was determined. Second, Cochrane's *Q* test was used to assess heterogeneity. Third, Hedges' *g* (Hedges, 1981; see also Cumming, 2012; Grissom & Kim, 2005; Hedges & Olkin, 1985), together with its standard error and its 95% CI, was computed as an effect size estimate to indicate the standardized mean difference. As explicated by Lakens (2013), it is preferable to report Hedges' *g* as an effect size estimate since it corrects for a slight positive bias inherent in Cohen's *d*. Finally, a *z*-statistic was calculated to test whether the combined effect size estimate is significantly different from zero.

Table 20 summarizes the model fits for all four meta-analyses.

Table 20. Model fit indices for the meta-analyses comparing modus ponens inferences and modus tollens inferences.

Model	Deviance	Log-likelihood	AIC	BIC	AICc
Endorsement Ratings					
Subjunctive	-0.55	0.28	3.45	0.84	15.45
Indicative	-1.34	0.67	2.66	0.04	14.66
Response Times					
Subjunctive	-0.57	0.28	3.43	0.82	15.43
Indicative	-0.84	0.42	3.16	0.55	15.16

Note. The random effects models were fitted using restricted maximum likelihood (REML).

Meta-Analysis 1. This meta-analysis estimated the effect in endorsement ratings between modus ponens inferences and modus tollens inferences when counterarguments were subjunctive. The effect size estimates were homogenous across the experiments, $Q(2) = 0.79$, $p = .675$. The combined effect size estimate amounted to $g = 0.69$ (95% CI [0.40, 0.98]), $SE = 0.15$. It significantly differed from zero, $z = 4.64$, $p < .001$. Hence, endorsement ratings were consistently and generally higher for modus ponens inferences as opposed to modus tollens inferences when subjunctive counterarguments were presented. Figure 23 shows the forest plot of meta-analysis 1.

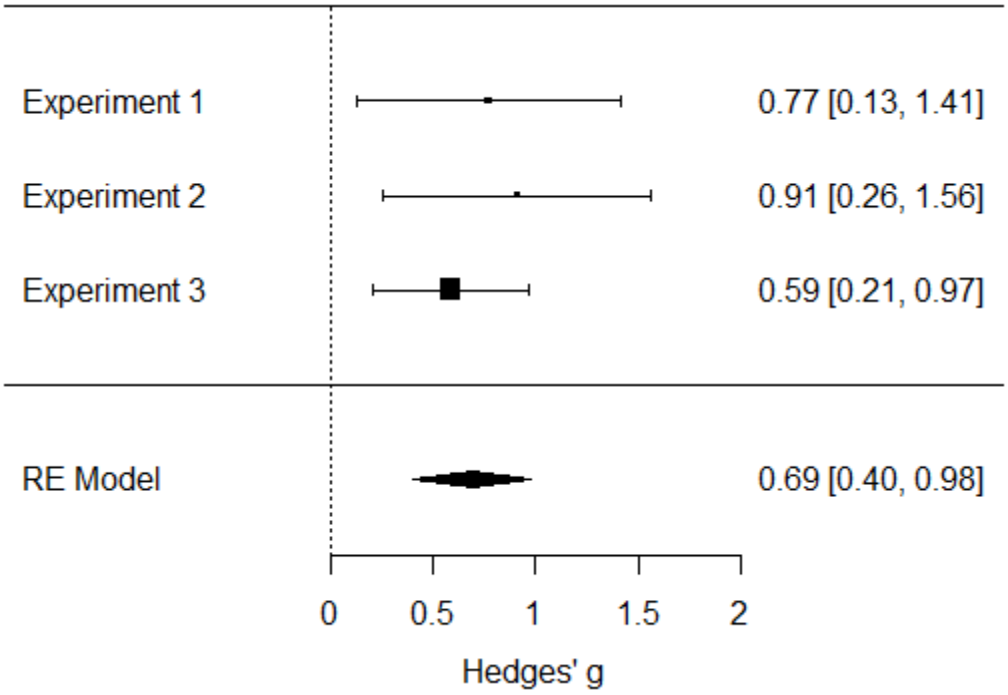

Figure 23. Forest plot of the effect size Hedges' *g* (and the 95% CI) in endorsement ratings between modus ponens inferences and modus tollens inferences when subjunctive counterarguments are presented.

Meta-Analysis 2. This meta-analysis estimated the effect in endorsement ratings between modus ponens inferences and modus tollens inferences when

counterarguments were indicative. The effect size estimates were homogenous across the experiments, $Q(2) = 0.05$, $p = .975$. The combined effect size estimate amounted to $g = -0.64$ (95% CI [-0.93, -0.35]), $SE = 0.15$. It significantly differed from zero, $z = -4.35$, $p < .001$. Hence, endorsement ratings were consistently and generally lower for modus ponens inferences as opposed to modus tollens inferences when indicative counterarguments were presented. Figure 24 shows the forest plot of meta-analysis 2.

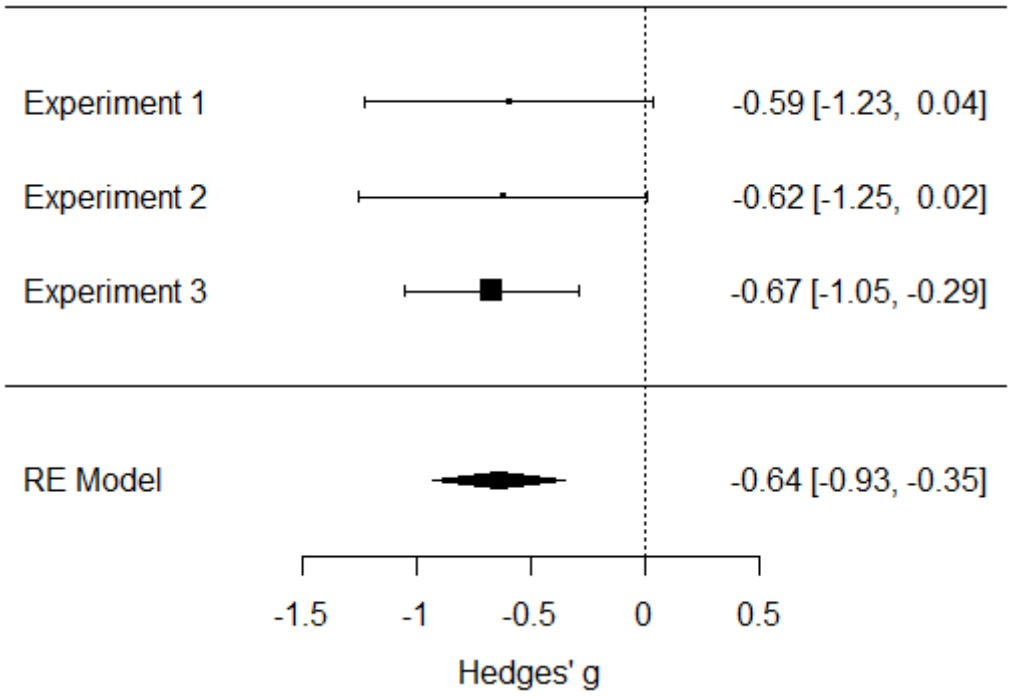

Figure 24. Forest plot of the effect size Hedges' *g* (and the 95% CI) in endorsement ratings between modus ponens inferences and modus tollens inferences when indicative counterarguments are presented.

Meta-Analysis 3. This meta-analysis estimated the effect in response times between modus ponens inferences and modus tollens inferences when counterarguments were subjunctive. The effect size estimates were homogenous across the experiments, $Q(2) = 0.89$, $p = .640$. The combined effect size estimate amounted to $g = -0.25$ (95% CI [-0.53, 0.04]), $SE = 0.15$. It differed from zero with

marginal significance, z = -1.70, p = .090. Hence, response times tended to be consistently and generally lower for modus ponens inferences as opposed to modus tollens inferences when subjunctive counterarguments were presented. Figure 25 shows the forest plot of meta-analysis 3.

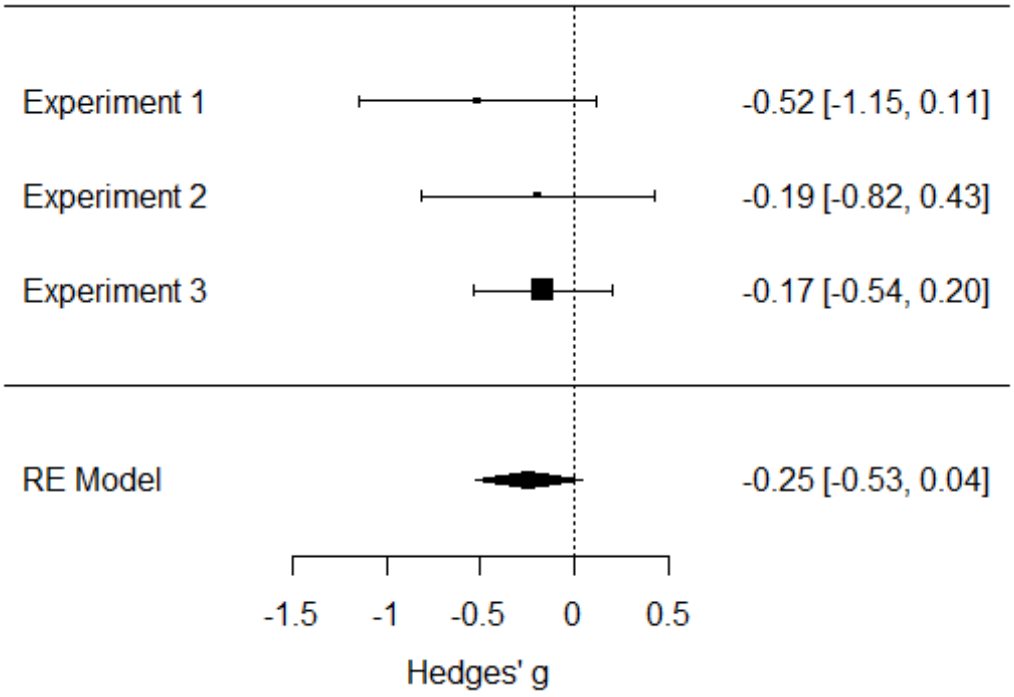

Figure 25. Forest plot of the effect size Hedges' *g* (and the 95% CI) in response times between modus ponens inferences and modus tollens inferences when subjunctive counterarguments are presented.

Meta-Analysis 4. This meta-analysis estimated the effect in response times between modus ponens inferences and modus tollens inferences when counterarguments were indicative. The effect size estimates were homogenous across the experiments, $Q(2) = 0.62$, $p = .735$. The combined effect size estimate amounted to g = -0.42 (95% CI [-0.71, -0.14]), SE = 0.15. It significantly differed from zero, z = -2.89, p = .004. Hence, response times were consistently and generally lower for modus

ponens inferences as opposed to modus tollens inferences when indicative counterarguments were presented. Figure 26 shows the forest plot of meta-analysis 4.

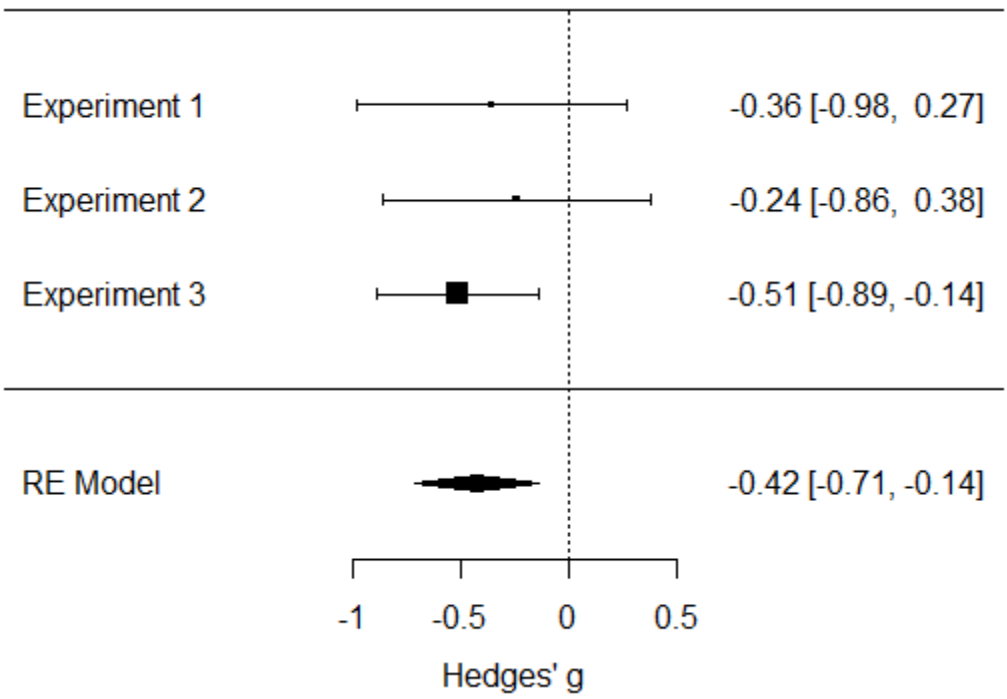

Figure 26. Forest plot of the effect size Hedges' *g* (and the 95% CI) in response times between modus ponens inferences and modus tollens inferences when indicative counterarguments are presented.

3.5.4 Discussion

The research aim of this chapter was to provide meta-analytic tests of the primary hypothesis permeating this thesis: Conclusion endorsement is higher for modus ponens inferences compared with modus tollens inferences when counterarguments are subjunctive, whereas it is lower when counterarguments are indicative. The guiding rationale was that if this interaction effect is in fact a consistent and generalizable psychological phenomenon, then the effect should meet two criteria: (1) the effect should appear homogenously across all experiments, and (2) the

combined effect size estimate should be significantly different from zero. As predicted, the meta-analyses revealed that these criteria are met. Meta-analysis 1 demonstrated that conclusion endorsement is higher for modus ponens inferences compared with modus tollens inferences when counterarguments are subjunctive. The effect was consistent across all experiments as evidenced by the test for heterogeneity. The combined effect size estimate was significantly larger than zero and displayed a substantial magnitude of the effect. Meta-analysis 2 demonstrated that conclusion endorsement is lower for modus ponens inferences compared with modus tollens inferences when counterarguments are indicative. Again, the effect was consistent across all experiments as evidenced by the test for heterogeneity. The combined effect size estimate was significantly smaller than zero and displayed a substantial magnitude of the effect. In addition, meta-analysis 3 and meta-analysis 4 speak for the tendency that modus ponens inferences require generally less time than modus tollens inferences, independent of their conjunction with a subjunctive or indicative counterargument, respectively. Although this effect regarding the response times appeared consistently across all experiments, please note that it reached merely marginal significance in meta-analysis 3, and the magnitudes of the combined effect size estimates for the response times were not as pronounced as for the endorsement ratings. Taken together, meta-analysis 1 and meta-analysis 2 provide robust and marked evidence for a medium effect concerning the interactive pattern in conclusion endorsement. In contrast, meta-analysis 3 and meta-analysis 4 only allow to speak of tentative support for the conjecture of a main effect of inference in response times. However, this finding should be interpreted cautiously and necessitates further primary research for its validation and before conclusive statements can be made.

I will now turn to a discussion of several methodological considerations. First, the conducted meta-analyses were based on a random effects model. Albeit I have

already outlined the advantages of a random effects model over a fixed effects model in the method section, another benefit should not be neglected: The conceptualization of a meta-analysis as a random effects model complies with standard scientific aims of generalization because a meta-analytic synthesis of research results enables to draw robust inferences that apply to a larger population of participants or a bigger universe of situations (Cooper et al., 2009). The realm of implications drawn from single studies is unavoidably restricted to a narrower epistemic scope. Second, the meta-analyses reported here are internal meta-analyses, meaning that they report experiments from the same series of studies conducted by the same investigator. An internal meta-analysis at the end of a research project is preferable to solely listing single studies since it can lend support for null findings, it can detect even small effects cumulatively, it helps to clarify the big picture and facilitates valid conclusions, it concisely summarizes findings and lightens the reader's cognitive burden, it counteracts an unhealthy and unsustainable culture of questionable research practices, it increases the comprehension of the study results, and it informs power calculations for subsequent primary research (Braver et al., 2014; Cumming, 2012, 2014; Goh et al., 2016; Maner, 2014). Third, a potential objection to conducting an internal meta-analysis might be the argument that a meta-analysis requires a certain number of primary studies to be conducted. The answer to this objection is: Yes, it needs more than one. As soon as a researcher conducted two (or more) studies, it is meaningful to perform a meta-analysis due to the plain and simple fact that a synthesis of two (or more) studies yields a more precise effect size estimate of the true effect than either study alone (Borenstein et al., 2021). Hence, two (or three as reported here) studies do already suffice to conduct an internal meta-analysis. The argument that one would need a daunting number of studies to legitimate a meta-analysis is simply false. This would be nothing more than an arbitrary convention without any logical or

statistical justification, enforced on researchers in a top-down fashion to fulfill some artificial guidelines that are the direct consequence of an ill-conceived cult of quantity maximization in science, which is doomed in the long run. I myself refuse to participate in that cult, which strikes me as nothing more than “an exercise in mega-silliness”, to adopt Eysenck’s (1978) words.

A valid concern with regards to meta-analysis is publication bias, which constitutes a formidable threat to the intellectual integrity of individual researchers and the scientific community as a whole. Rosenthal (1979) has already described this issue as the so-called file drawer problem. In the same vein, other authors have argued and shown that on average published studies display larger effect size estimates than unpublished studies (e.g., Lipsey & Wilson, 1993; McNemar, 1960; Smith, 1980; Spellman, 2012). A career-driven preferential publishing of significant results, and the storage of non-significant results in a researcher’s file drawer, leads to a systematic overestimation of combined effect size estimates when accumulating evidence over several primary studies in a meta-analysis. The severe implications of publication bias for the detriment of science are undeniable (Ioannidis, 2005). I see two lines of action to counteract publication bias: On the one hand, junior researchers should include all studies of a research project when performing a meta-analysis, regardless of the *p*-values of individual studies. Moreover, they should have the courage to do so even in the face of being punished for it by power holders in the academic system who have not yet come to understand the damage they cause by punishing high-quality work and rewarding high-quantity work. On the other hand, senior researchers may rethink their funding decisions, and editors might select contributions that showcase excellent exemplars to serve as a prototype and inspiration for others. This way, junior and senior researchers can jointly create an intellectual environment that successfully fosters excellent scholarship based on strong inference (Platt, 1964).

Lastly, I wish to emphasize that a meta-analysis does not need to be the final end of a research line. A meta-analysis that concludes that a phenomenon is completely understood and no future research is needed should be viewed with extreme skepticism (Cooper et al., 2009). For instance, if a researcher derives a novel hypothesis from the findings synthesized in a meta-analysis, or even develops a new theory based on a meta-analysis, then new primary research is needed to test it. Put differently, if meta-analytic data are used to derive a hypothesis, then new data must be used to test it—the generation of the hypothesis and the data for testing it must be independent (Wachter & Straf, 1990). The combined effect size estimate of a meta-analysis can then be used for the power calculations and sample size planning of such new primary research. Consequently, the determination of a required sample size for primary research becomes more solid (Fiedler et al., 2012; Lakens, 2013; Maxwell et al., 2008).

4 General Discussion

The general discussion section of the present thesis aims to summarize and interpret the findings from the empirical evidence section. To this end, the general discussion is divided into four parts: First, I synthesize the research results I produced in a joint review of all three conducted experiments, the mixed model analyses, and the meta-analyses. Second, I discuss relevant implications of my findings for theory, related concepts, methodology, and applications. Third, I reflect on limitations in order to abduce conjectured explanations thereof, which can be used to inspire future research. I provide a concise outline for possible future directions that seem promising to me. And fourth, I conclude my thesis with some final codas.

4.1 Research Synthesis

The central research objective of the present thesis was to study the pragmatic

modulation of rational argumentation in conditional reasoning with counterarguments. For this purpose, I conducted three experiments, reanalyzed the results using mixed models, and computed meta-analyses.

Experiment 1 served as the first original test of the deduced predictions. In line with my a priori theorizing, I obtained results showing that, overall, modus ponens conclusions are endorsed more frequently than modus tollens conclusions. I also found that, overall, counterarguments reduce conclusion endorsement. Most importantly, the findings confirmed my prediction of an interaction between the inference type of the conditional (modus ponens versus modus tollens) and the linguistic mode of the counterargument (subjunctive versus indicative). Specifically, as hypothesized, conclusion endorsement was higher for modus ponens compared with modus tollens when counterarguments were subjunctive. Conversely, conclusion endorsement was higher for modus tollens compared with modus ponens when counterarguments were indicative.

Experiment 2 served as a direct replication. *Ceteris paribus*, the only aspects that changed were the nationality and the native language of the participant sample. Hence, all study materials (i.e., instructions, stimuli, and dependent measures) were translated to Italian using the method of back-translation (Brislin, 1970). As expected, the findings of the previous experiment were replicated. Notably, the response pattern remained robust, providing suggestive evidence for its invariance across different language-culture groups. Thereby, the successful replication increases the confidence in the objectivity of the operationalization of the target constructs into the experimental procedures, raises the confidence in the reliability of the employed measures, enhances the confidence in the validity of the results, and elevates the confidence in the robustness and width of scope of the evidential response pattern.

Experiment 3 served as a second replication and an extension of the former

experiments. Importantly, I implemented a nested experimental design that incorporated and tested the relevance of counterarguments as an essential boundary condition for the findings. Again, the general findings from the previous two experiments were replicated. Importantly, however, the interaction between the inference type of the conditional and the linguistic mode of the counterargument was moderated by relevance. As hypothesized, the nuanced interaction only reemerged under the boundary condition of relevance. If this condition was not met, then the interplay of inference type and linguistic mode regressed to two simple main effects. This suggests that relevance functions as a pivotal catalyst of the significant interaction effect between a conditional's inference type and a counterargument's linguistic mode, affecting conclusion endorsement in a multiplicative fashion.

The mixed model analyses served as a reanalysis of the findings from all three experiments, taking into account the multilevel structure of the data. That is, data were not analyzed in an aggregated fashion. Instead, data were analyzed in an unaggregated trial-by-trial fashion, which prevents information loss by modeling the random variation between participants. In line with my prediction, results remained virtually unaffected by including the participant variable as a random intercept into the individual model terms. Most crucially, the interaction between inference type and linguistic mode—as well as its modulation by relevance—remained stable. This indicates that the findings are not an artefact of my data analysis strategy. In fact, my findings are demonstrably reproducible across various data analysis protocols.

The meta-analyses served to demonstrate the homogeneity and robustness of the major finding (namely, the interaction between inference type and linguistic mode) across all conducted experiments, and estimated its combined effect size. As expected, the meta-analyses revealed that the interaction effects of inference type and linguistic mode were highly homogenous across the three experiments. Robustly,

conclusion endorsement was higher for modus ponens as opposed to modus tollens when counterarguments were subjunctive, with a combined effect size estimate of Hedges' *g* = 0.69. Robustly, conclusion endorsement was lower for modus ponens as opposed to modus tollens when counterarguments were indicative, with a combined effect size estimate of Hedges' *g* = -0.64. Given the absolute magnitudes of these meta-analytically derived effect sizes, I consider the predicted interaction as a conclusive finding, which demands its preparation for publication in a top-tier, peer-reviewed international journal. An auxiliary finding of the meta-analyses indicates that response time was faster for modus ponens as opposed to modus tollens when counterarguments were subjunctive, with a combined effect size estimate of Hedges' *g* = -0.25. Likewise, response time was faster for modus ponens as opposed to modus tollens when counterarguments were indicative, with a combined effect size estimate of Hedges' *g* = -0.42. This suggests that modus ponens inferences require less processing time than modus tollens inferences, both for subjunctive and indicative counterarguments.

A proximate interpretation of the evidence summarized above regards the idea that pragmatics (Grice, 1975) stands the test of time when it comes to identifying sustainable overarching frameworks for the scientific study of rational argumentation. Since rational argumentation can be operationalized and measured via conditional reasoning with counterarguments, and conditional reasoning with counterarguments has been demonstrated to be decisively influenced by pragmatic factors, it follows that the experimental variation of pragmatic factors in studies of conditional reasoning with counterarguments can help us to make novel inductive inferences with respect to the nature of rational argumentation. Indeed, this reasoning is congruent with several empirical findings demonstrating the marked role of pragmatics for conditional reasoning (Douven et al., 2020; Espino et al., 2020; Evans, 1983; Gazzo Castañeda

& Knauff, 2016b, 2019, 2021c; Hilton et al., 1990; Khemlani et al., 2018; Roberge, 1978; Thompson, 1994, 1995; Valiña et al., 1999). Especially, the present findings advance theory development in the study of conditional reasoning because they identify specific pragmatic variables (e.g., counterarguments, linguistic mode, relevance) that directly exert measurable variations in the response patterns of behavioral data, which, in turn, allow for inductions towards the validity of the theoretical assumptions that the respective findings rest upon. This, consequently, helps to find and correct problems in theories of conditional reasoning. In particular, it is constructive in that it directs attention towards open questions that are yet to be addressed (Byrne & Johnson-Laird, 2009), and aids in developing a more comprehensive theory of conditional reasoning by acknowledging the roles of meaning and pragmatics in human inference (Johnson-Laird & Byrne, 2002). On a related note, my findings highlight the need of the psychology of reasoning to bring more concepts into the picture and not solely focus on logical form and the dictates of classical logic. Such other concepts may encompass persuasion, argument, probability, Bayes' theorem, world knowledge, and non-monotonicity, to name but a few. These deliberations are consistent with the claims of the so-called new paradigm psychology of reasoning (Oaksford & Chater, 2020). However, please bear in mind that the alleged new paradigm might not be a new paradigm after all, because it can just as well be read as a natural continuation of the so-called old paradigm; plus, the notion that there exists only one paradigm in the psychology of reasoning is arguably too reductionist anyway (Knauff & Gazzo Castañeda, 2021). It also clarifies the importance of careful nomenclature, and thus helps to avoid jingle (i.e., the same name for different constructs) and jangle (i.e., different names for the same construct) fallacies.

The basic finding of a facilitated processing of modus ponens inferences in comparison to modus tollens inferences, as evidenced both in conclusion

endorsements and response times, can be gauged in the light of three viable theoretical accounts—*formal rules*, *suppositions*, and *mental models*—, all three of which pose a different explanation to interpret the processing advantage of modus ponens over modus tollens. *Formal rule theories* (Braine & O'Brien, 1998; O'Brien, 2009; Rips, 1994) postulate that conditional reasoning follows a set of formal inference rules akin to a logical calculus. These accounts explain the processing advantage of modus ponens over modus tollens with the existence of a formal rule of inference for the affirmative modus ponens inference in the human mind, while a comparable formal rule for the negative modus tollens inference is not implemented in the human mind. *Suppositional theories* (Evans, 2007; Evans & Over, 2004; Evans et al., 2005), being based on the Ramsey test (Ramsey, 1929/1990), postulate that conditionals elicit suppositional thinking, meaning that reasoners suppose the conditional's truth and think about its consequences. The supposition represents the case needed for the affirmative modus ponens inference, but not the case needed for the negative modus tollens inference. *Mental model theory* (Johnson-Laird, 1983, 2006; Johnson-Laird & Byrne, 2002) postulates that conditional reasoning depends on the construction, inspection, and variation of possibilities. According to this account, mental models represent the possibility needed for the affirmative modus ponens inference, but not for the negative modus tollens inference. Notably, the mental model theory can also be read as a dual-process account of conditional reasoning, postulating an intuitive process underlying the modus ponens inference and a deliberative process underlying the modus tollens inference. To explain, consider the following conditional:

If she plays the piano, then music fills the room.

During the intuitive process, only a single mental model is constructed explicitly, representing the possibility in which both clauses (antecedent and consequent) hold. Although it is realized that the conditional rule might be false, these possibilities are not represented explicitly. Therefore, the mental model looks as follows:

piano *music*

…

The first line of the mental model (i.e., piano and music) represents the explicit possibility of the model. The second line of the mental model (i.e., the ellipsis) represents other possibilities, which are not yet fleshed out and therefore only implicitly represented. This intuitive process suffices to draw the easy modus ponens inference. However, it does not suffice to draw the modus tollens inference. This necessitates the deliberative process, which relies on working memory. During the deliberative process, all three possibilities, which refer to the specific combinations of truth values of the first premise and the second premise that yield true implications according to propositional logic (Knauff & Knoblich, 2017; see Table 21), are now envisaged from implicit into fully fleshed out, explicit possibilities. Therefore, the mental model looks as follows:

piano *music*

¬piano *music*

¬piano *¬music*

The explicit representations of these three possibilities suffice for the difficult negative modus tollens inference, albeit demanding more working memory capacity.

Table 21. Truth table of propositional logic, adapted from Knauff and Knoblich (2017, p. 537).

	Premise 1	Premise 2	Conjunction	Disjunction	Implication	Biconditional
Natural language	P	Q	P and Q	P or Q	if P then Q	if and only if P then Q
Symbolic form	P	Q	$P \wedge Q$	$P \vee Q$	$P \rightarrow Q$	$P \leftrightarrow Q$
Truth value	T	T	T	T	T	T
	T	F	F	T	F	F
	F	T	F	T	T	F
	F	F	F	F	T	T

Note. T = true; F = false.

Hence, to conclude, the dual-process account of mental model theory affords an elegant and parsimonious theoretical explanation for the facilitated processing of modus ponens compared with modus tollens.

Essentially, the interaction effect of inference type of the conditional and linguistic mode of the counterargument offers the most stimulating opportunity for theorizing proximal interpretations. The nuanced processing of modus ponens versus modus tollens as a function of subjunctive versus indicative counterarguments might have to do with the idea that subjunctive counterarguments may be construed as counterfactual (or at least hypothetical) incidents, whereas indicative counterarguments may be processed as factual incidents. Notwithstanding counterfactuals are not uncommon in everyday discourse (Byrne, 2002, 2005; Mandel et al., 2005; Markman et al., 2008), they differ in their meaning and inferential consequences from factual propositions (Byrne & Johnson-Laird, 2009). Factual propositions only represent real states of affairs. However, counterfactual propositions can represent both a real incident as well as its negation. In other words, a factual proposition leaves room for one possibility; a counterfactual proposition, by contrast, leaves space for two possibilities (Bennett, 2003; Williamson, 2007). Therefore, a factual proposition primes the construal of the factual possibility only, while a counterfactual proposition primes the construal of both the factual and the counterfactual possibilities (De Vega et al., 2007; Santamaria et al., 2005). People endorse modus tollens more frequently when the conditional is interpreted as counterfactual instead of factual (Byrne, 2005, 2007; Byrne & Johnson-Laird, 2009). A factual indicative counterargument rather than a counterfactual subjunctive counterargument has arguably more power in rendering the modus tollens conditional itself counterfactual. This could explain the—at first glance counterintuitive but theoretically well justified—finding that modus tollens endorsement is more

pronounced when indicative counterarguments are presented compared with subjunctive counterarguments. However, I want to emphasize that no theory has a monopoly on truth. Still, I consider the mental model theory a parsimonious and elegant proximate approach because it explains many aspects of the observed findings well and in a straightforward fashion.

Importantly, furthermore, the robust interaction of inference type and linguistic mode is consistent with the prediction I derived from the mismatch principle (Elio & Pelletier, 1997; Hasson & Johnson-Laird, 2003; Johnson-Laird et al., 2004; Revlin et al., 2001). However, it should be noted that Byrne and Walsh (2002) report results that seem to contradict the mismatch principle. This contradictory finding may have occurred because they utilized a fundamentally different experimental setup. In their design, participants drew a first explicit conclusion *before* the inconsistent fact (i.e., the counterargument) was presented. This procedure biases participants to perceive the inconsistency between the conclusion and the counterargument, and not between the default premises (i.e., major premise and minor premise) and the counterargument. This difference in experimental design might explain the incongruent finding of Byrne and Walsh (2002). As a matter of fact, Hasson and Walsh (2003) compared the two procedures and were able to replicate both sets of results. This nicely demonstrates that complex processes like human inference need not be investigated using a single standard approach. I agree with Dhami et al. (2004) that not every researcher should stick to the same standardized design only for the sake of avoiding conflicting results. If anything, a researcher should be clear about the goal of the study from the outset, clearly state the hypothesis, and specify the context to which generalization is intended. This helps prevent methodological monopolization and counters pluralistic ignorance, both of which would otherwise impede scientific progress. Instead—in the spirit of Paul Feyerabend (1975, 1978)—epistemological plurality is a much better

choice! It enriches the diversity of psychological science and advances our understanding of human inference.

Considering this synopsis of all findings of my thesis, the present synthesis can also be seen as a starting point for a more comprehensive and integrated model of rational argumentation—a model designated to describe, explain, and predict the tight interplay of the cognitive mechanisms and the pragmatic constraints of rational argument. It goes without saying that this is a task that can demand a whole scientific career. Nonetheless, I believe the present synthesis could provide a first aspect for such an enterprise to gain momentum. Figure 27 draws a rough sketch of a cognitive-pragmatic interface model of rational argumentation.

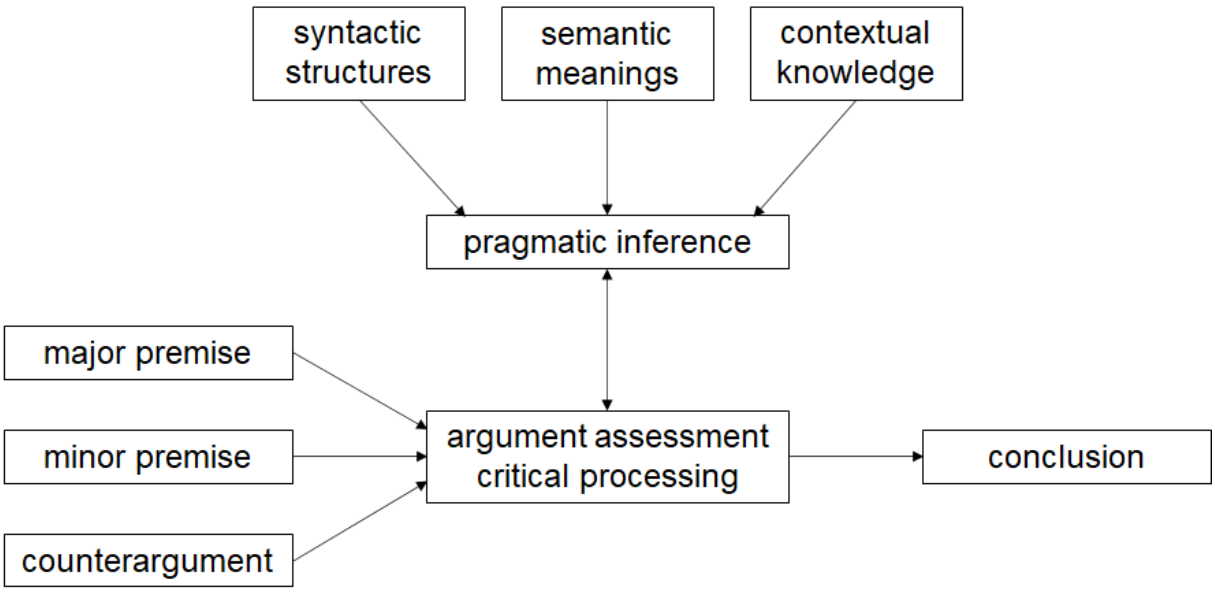

Figure 27. Blueprint of the cognitive-pragmatic interface model of rational argumentation, adapted from Oswald (2007, p. 193).

Needless to say, this first outline is tentative; it requires further refinement. A more fine-grained specification of the model can be reached by means of mathematical formalization, computational modeling, and continued experimental testing. Note that

this model is, in part, inspired by the pragma-dialectical and relevance-theoretic interfacing model proposed by Oswald (2007).

Taken together, the research synthesis has bundled the findings of Experiment 1, Experiment 2, Experiment 3, the mixed-model analysis, and the meta-analysis, offered proximal interpretations of the findings based on basal cognitive and pragmatic principles, and provided a first blueprint of a cognitive-pragmatic interface model of rational argumentation that may motivate future research. Next, I intend to draw more distal implications for theory, related concepts, methodology, and applied contexts.

4.2 Implications

This chapter continues to interpret the obtained findings. Here, I will unpack the interpretative space by consulting the extant literature and by reflecting on meta-theoretical issues of the argumentative theory of reasoning, relevance theory, and related concepts whose inspection may be fruitful for future theory development and research in rationality and argumentation. Furthermore, I will discuss methodological and practical implications that follow from the insights gained on the theoretical and the conceptual level.

4.2.1 Theoretical Implications

“There is nothing as practical as a good theory” (Lewin, 1943, p. 118). Kurt Lewin’s quote is probably one of the most often cited aphorisms in psychology. In this spirit, for the accrued evidence to reach its full applied potential, it is key to reflect on the theoretical implications and the broader meta-theoretical consequences thereof. Science ought to progress by finding and correcting problems in theories, thereby fostering an incremental increase in a theory’s adequacy. Philosophers of science have given researchers the criteria against which to measure the success (or failure) of a theory. A good scientific theory makes testable predictions that are falsifiable (Popper, 1935, 1963, 1972). Also, a good scientific theory consists of a hard core of

main assumptions and a protective belt of auxiliary assumptions (Lakatos, 1970). Further meta-theoretical criteria of theory evaluation include conservatism, parsimony, generality, refutability, and precision (e.g., see Gawronski & Bodenhausen, 2015). These are the fundamental normative standards to be considered in theory evaluation and theory development alike. Moreover, contrary to the popular mainstream opinion that the low credibility (e.g., replicability problems) of psychological science can be tackled with a unilateral focus on method-based solutions and "open science" regimentation (instead of encouraging creativity and rewarding professional ethos), strong theorizing and a greater emphasis on theory criticism by argument might be a better way forward (e.g., see Fiedler, 2017, 2018; Gray, 2017; Oberauer & Lewandowsky, 2019; Szollosi & Donkin, 2021). Besides these obvious epistemic advantages of a theory-driven approach, another beneficial side-effect is nicely summarized by David Trafimow (2015):

> Had researchers taken a philosophical perspective, much of the research need not have been performed. My hope is that the graduate student or professor who wishes to conduct research will consider the larger philosophical and conceptual issues before embarking on what might be an expensive and time-consuming excursion that leads nowhere. (p. 264)

The quote points to the fact that investing time in theory development might seem effortful and demanding in the beginning—and it is, to be frank. However, ultimately, making this investment will prove to be a more effective and sustainable strategy in the long run. I argue that this should be a continuous practice because theorizing is productive both before and after testing: before testing because it helps to clearly state the research objectives and specify the hypotheses; and after testing

because drawing wider theoretical implications helps to identify and disseminate which parts of a theory must be refined and what other hypotheses are worthwhile for follow-up testing. In the following, I discuss these wider theoretical implications. I concentrate on two questions: (1) what are the theoretical implications for the argumentative theory of reasoning; and (2) what are the theoretical implications for relevance theory?

What are the theoretical implications for the argumentative theory of reasoning? First, a major finding of the present thesis is that people are quite proficient in identifying argument quality and, most importantly, they act rationally because they filter the essential (counter-)arguments and virtually disregard the ones that are inconsequential. Moreover, they also draw fine-grained and differential conclusions, not only based on the consequentialist effects of counterarguments, but they also integrate the inference type of a conditional rule and subtle linguistic cues, like a counterargument's mode, in a highly sophisticated and nuanced inference process. These findings support the argumentative theory of reasoning (Mercier, 2016a; Mercier & Sperber, 2011, 2017) since they are consistent with a core assumption of the theory, namely that humans show good performance during argument evaluation. They are highly engaged in demanding quality control processes during the evaluation of arguments and counterarguments, and accept only (counter-)arguments if they are strong. Hence, the accumulated evidence corroborates a central theoretical layer of the argumentative theory of reasoning. Johnson (2011) claims that these implications are independent of level of expertise; therefore, they should apply to professional thinkers and arguers (e.g., scientists and philosophers), too. The confirmative evidence reported in the present thesis does also contribute to the establishment of an integral theory of the psychology of argumentation, because it marries cognitive and pragmatic insights from different disciplinary strands (Hornikx & Hahn, 2012). The view I adopted in the present thesis bridges these different strands and shows how closely they are

intertwined in reality. Note that the theoretical implications for the argumentative theory of reasoning specifically refer to argument evaluation. Participants in the conducted experiments were prompted to evaluate the conclusiveness of arguments, not to produce arguments themselves. As for the production of arguments, one would expect quite a different pattern, namely that people are strongly biased towards producing arguments that confirm their own beliefs. In this sense, another theoretical implication lies in the notion that the confirmation bias (Nickerson, 1998), as dominant as it might be during argument production, has its boundary conditions. During argument evaluation, people actually assess the quality of arguments and counterarguments rather rationally and unbiased. Even the difficult modus tollens inference is being made more often when people are asked to evaluate an argument instead of being instructed to pass a reasoning task (Thompson et al., 2005). The evidence reported in my thesis converges with the results from Thompson et al. (2005). Although, on average, the relative performance for modus tollens is still lower compared with modus ponens, a closer look at the absolute scale values for modus tollens indicates that it is in fact endorsed more strongly than is often suggested. Moreover, in the specific case of presenting indicative (as opposed to subjunctive) counterarguments, modus tollens endorsement is even higher than modus ponens endorsement. This suggests that embedding reasoning tasks in argumentative contexts (e.g., via instruction, imagery, group discussion, etc.) improves reasoning performance and facilitates valid inferences during argument evaluation. Pennington and Hastie (1993) found that, given an argumentative context, valid modus tollens inferences are surprisingly common even during argument production. A further aspect of the current thesis is that my findings illuminate the crucial value of counterarguments for rational argumentation. They imply that the availability of arguments that challenge a reasoner's initial perspective can be beneficial for drawing strong inferences, which ultimately yields

rational decisions. Three specific conditions are key for preparing a situation that optimizes the benefits of counterarguments: the exchange of conflicting arguments (Mercier & Landemore, 2012; Thompson, 2008), the evaluation of counterarguments (Mercier, 2011b; Mercier & Landemore, 2012), and a general motivation to be truth-oriented (Mercier & Sperber, 2011). When these conditions are met, counterexamples will exert their full power. On a meta-theoretical level, the empirical findings of this thesis suggest that argumentative context (e.g., provided by counterarguments) improves rational thought. This insight is in stark contrast to a purely individualistic view on rationality, which is prevalent at least since René Descartes' (1637) famous *je pense, donc je suis* (in Latin: cogito, ergo sum). Many psychologists have adopted this purely individualistic perspective (e.g., Kahneman, 2003; Stanovich, 2004). My thesis represents an alternative view: The main function of human rationality is social. In essence, human reasoning is adapted to produce good arguments in order to convince others, and to evaluate the arguments of others in order to develop more rational beliefs of the world. I do not see this theory as a rival. Instead, let me stress that I do consider this social view of rationality as *complementary* to the individualistic view. Both views are important in increasing our understanding of human rationality. Of course, the argumentative theory of reasoning (Mercier & Sperber, 2011, 2017) is the current main proponent of the social view of reasoning. However, other theoretical accounts in support of the social view have been documented (Baumeister & Masicampo, 2010; Billig, 1996; Gibbard, 1990). Before closing this paragraph on the theoretical implications for the argumentative theory of reasoning, I want to point out some boundary conditions that should not be neglected. Hahn and Collins (2021) claim that interindividual variation that is due to different education levels (Kuhn, 1991) or epistemological beliefs (Kuhn et al., 2000; Kuhn et al., 2010) might, at least to some degree, affect people's rational-argumentative faculties. Hahn and Collins (2021)

define two prototypes of argumentative thinkers, which they label “multiplists” and “evaluativists”. These two types have distinct characteristics. “Mutliplists” regard facts as freely chosen opinions. They tend to refrain from engaging in argumentation altogether, if they can. “Evaluativists”, on the other side, view facts as judgments that can be assessed against normative standards of evidence. They are more probably keen to engage in argumentation. If the hypothesis of two distinctive types of argumentative thinkers is valid, then one should predict and be able to demonstrate different response patterns in studies on conditional reasoning with counterarguments. Therefore, future studies could test this hypothesis, either quasi-experimentally with naturally occurring groups of “multiplists” and “evaluativists”, or by experimentally priming a “multiplist”-focus versus an “evaluativist”-focus. A final boundary condition I would like to draw concerns the scope of the argumentative theory of reasoning. The theory defines (argumentative) reasoning as a specific meta-representational mechanism (Mercier, 2013a). Therefore, the theory specifically refers to reflective system 2 inferences. Albeit it also gives an account of the functions of intuitive system 1 inferences, its main focus—especially with respect to argument evaluation—lies on reflective inferences driven by system 2 (Mercier & Sperber, 2009).

What are the theoretical implications for relevance theory? A central insight gained from the work reported in the present thesis is that people are apt to monitor the quality of counterarguments as a function of relevance. Counterarguments that are relevant since they match the argumentative context need more processing time than counterarguments that are irrelevant since they do not fit the argumentative context. Furthermore, the nuanced interaction effect between inference type and linguistic mode is only present when counterarguments are relevant. In the case of irrelevant counterarguments, this interaction breaks down and regresses back to two simple main effects. These findings imply that people utilize relevance as a central criterion

for the evaluation of (counter-)arguments during rational argumentation. Hence, the present findings corroborate a core assumption of relevance theory—the cognitive principle of relevance, stating that human cognition can be characterized by a tendency to maximize relevance (Sperber & Wilson, 1995). My results confirm the relevance-theoretic notion that an input (i.e., a counterargument) is considered relevant to the degree that it combines with contextual information in order to achieve a useful cognitive output. The findings also speak for the prediction that relevance maximization is realized by virtue of maximizing positive cognitive effects whilst keeping the required processing effort reasonably low (Sperber & Wilson, 1986, 1995). These preconditions—maximization of positive cognitive effects and minimization of processing effort—were only met in the condition in which relevant counterarguments were presented. The condition with irrelevant counterarguments did not allow for an elicitation of cognitive effects that would have increased a pragmatically reasonable and meaningful inference, nor would it have led to a reasonably low processing effort because the mental effort to integrate irrelevant counterarguments into the argumentative context would have been high. In exquisite detail, integrating irrelevant counterarguments would be much more costly because it necessitates much more mental resources to represent the input, access the argumentative context, and derive the respective cognitive effects. Therefore, people tend to disregard irrelevant counterarguments and rather direct their attention and resources towards relevant counterarguments, because the latter are the ones that keep the ratio of positive cognitive effects and processing effort optimal. A second interpretation for the disregard of irrelevant counterarguments is embedded in the communicative principle of relevance, stating that every overt communicative act conveys a presumption optimal relevance for what is communicated itself (Sperber & Wilson, 1995). Obviously, the irrelevant counterarguments were incongruent with the presumption of optimal

relevance. This led to a violation of the communicative principle of relevance, which, consequently, resulted in the virtual neglect of counterarguments that were irrelevant to the argumentative context at hand. Put differently, participants evaluated irrelevant counterarguments as not being worth deeper processing, whereas relevant counterarguments were the ones considered most compatible with the specific argumentative context. Taken together, I can therefore conclude that the findings of my thesis support relevance theory (Sperber & Wilson, 1986, 1995) in that they (1) corroborate that relevance follows a function of maximizing positive cognitive effect and minimizing mental processing effort, (2) confirm the cognitive principle of relevance that human cognition tends to be geared to the maximization of relevance, and (3) speak for the communicative principle of relevance that every act of overt communication conveys a presumption of its own optimal relevance. However, while I did find evidence for these major relevance-theoretic core assumptions on a functional level, there are still much unanswered questions concerning the underlying mechanistic processes. Therefore, I consider it highly fruitful to shed a light on such questions in future studies by addressing the algorithmic level of analysis. For example, it seems worthwhile to me to invest effort in investigating the relevance-guided comprehension heuristic as well as the subtasks involved in this comprehension process (Sperber et al., 1995; Wilson & Sperber, 2012a). A related research question would be: How do people *interpret* an argument? This would involve examining how people resolve ambiguities and referential indeterminacies, how they go beyond linguistic meaning, how they supply contextual assumptions, how they compute implicatures, how they stop information sampling when the expected level of relevance is achieved, how they construct an appropriate hypothesis about explicatures, how they construct an appropriate hypothesis about the intended contextual assumptions (i.e., the implicated premises), and how they construct an appropriate hypothesis about

the intended contextual implications (i.e., the implicated conclusions), to name but a few. I am convinced that asking these how-questions will tremendously advance our understanding of rational argumentation. Another theoretical implication of the current findings for relevance theory concerns the validity and the adequacy of the theory in and of itself. My work shows that relevance theory is a good theory in that it allows for the deduction of empirically testable a priori hypotheses, which, in turn, can be translated into explicit predictions of outcomes in experimental tests. Hence, relevance theory meets two gold standards of a strong theory—it is testable and falsifiable. Also, the results of my confirmatory hypothesis tests indicate that relevance theory is hard to vary. Nevertheless, this is not to say that researchers should stop testing the theory. To the contrary, we should continue with attempting to falsify the theory (or some of its assumptions). This is the game of science. And it is this game of science that will show whether or not relevance theory stands the test of time. Hitherto, relevance theory remains a successful post-Gricean theory of rational argumentation. Aside from my own findings, an abundance of theorizing and research reinforces this conclusion (e.g., Belligh & Willems, 2021; Blakemore, 2001; Franken, 1997; Gibbs Jr. & Tendahl, 2006; Huang & Yang, 2014; Levinson, 1989; Medin et al., 2003; Nicolle, 1998; Sperber & Wilson, 1997; Wearing, 2015; Yuan et al., 2019; Yus, 1998, 2003, 2010). Eventually, the findings of this thesis imply that relevance theory may serve as a comprehensive meta-level theory of information processing during argument evaluation. It is consistent with other cognitive theories of information processing, such as mental model theory (Johnson-Laird, 1983) and Fodor's (1983) modularity hypothesis, stating that human information processing is modulated by the desire to achieve successful outcomes (i.e., high positive cognitive effects), and by the tendency to do so as efficiently as possible (i.e., low processing effort). In this sense, relevance theory draws a much more positive picture of human rationality than some more recent theories (e.g., Haidt,

2001), which try to portray humans as instinct-driven, emotional animals that base their decisions on intuition and that only use reasoning as a post hoc construction in order to rationalize their irrational choices. Of course, swimming along with what seems *en vogue* on the academic market appears attractive to opportunistic people. But appearances are deceitful. I, instead, have shown that we can be much more optimistic when it comes to assessing rationality, especially in argumentative contexts. People are rational arguers who produce strong arguments to convince others, and who successfully evaluate arguments as a function of relevance in order to distinguish between strong arguments and weak ones. I admit, though, that this positive conceptualization of human rationality is currently unpopular and therefore its advocacy requires some *chutzpah*. I, for my part, consider it a risk worth taking!

4.2.2 Conceptual Implications

Whereas the last chapter illuminated the theoretical implications of this thesis, the present chapter reflects on some conceptual implications that should be taken into account. Reflecting on both inherent concepts of a theory as well as related concepts is important for various reasons. First, precise and consistent nomenclature of the concepts that scientists use enables communication among colleagues and with the general public to be most effective. Conceptual terminologies should be unequivocal in order to avoid jingle (i.e., the same name for different concepts) and jangle (i.e., different names for the same concept). Second, thinking conceptually may aid the construction of nomological networks (Preckel & Brunner, 2017). Nomological networks are multi-layered representations of the latent concepts (i.e., constructs) of a study and their manifest, empirically measurable variables (i.e., observations). Specified linkages and correspondence rules indicate the interrelations among concepts, among variables, and between concepts and variables. Thereby, nomological networks are extremely helpful for theory development and the deduction

of novel hypotheses (Alavi et al., 2018). They further clarify ways forward with respect to how a construct is meaningfully operationalized. And lastly, nomological networks assist the establishment and evaluation of construct validity by assessing the degree to which a construct operates as it should within a system of related constructs (Cronbach & Meehl, 1955; Liu et al., 2012; Shadish et al., 2002). Similar constructs should yield highly positive correlations and thus indicate convergent construct validity. By contrast, opposite constructs are expected to yield highly negative correlations and consequently indicate divergent construct validity. Unrelated constructs should yield correlations of or approximate to zero, which suggests a lack of construct validity. These conceptual considerations are crucial to think about where conceptual overlaps or family resemblances can inspire scientific innovation, where it produces contamination and confusion, or where it simply is redundant or obsolete. I will now continue with elucidating some conceptual issues that may have implications for the interpretation of my findings that should not be ignored.

One conceptual distinction, I argue, that should be made explicit is the distinction between the concepts *reason* and *inference*. Both words frequently appear in this thesis. Therefore, it is of utmost importance to clearly demarcate them on the conceptual level to allow for the facilitation of the right implications as well as the avoidance of false implications. Reason is a functional unit that operates on the computational level. It primarily represents the input to a cognitive unit. It can also reflect the goal(s) of a cognitive computation. Therefore, in the case of the present thesis, reason mainly refers to the to be computed propositions, namely premises (major premises and minor premises), arguments (as compounds of premises), and counterarguments (counterfactual premises and contradictory premises). Note that the concept also encompasses the conclusions since they are the computational outputs and can therefore be regarded as functions of reasons. Inference, on the other side, is

a mechanistic unit that operates on the algorithmic level. It represents the underlying processes upon which a cognitive unit operates. Therefore, in the present case, inference describes the mediating principles that lead from input to output, that is, from arguments and counterarguments to conclusion. In summary, it is therefore important to keep in mind that all functional implications I drew specifically refer to arguments, counterarguments, and conclusions as propositional entities, whereas all mechanistic implications I drew must be interpreted on the level of the cognitive processing itself. In this, my reasoning concurs with the authors of relevance theory and the argumentative theory of reasoning, who have repeatedly emphasized that the concepts reason and inference should be used carefully in the interpretation of research findings embedded in their theories (Mercier, 2016a; Mercier & Sperber, 2011).

Another conceptual differentiation, which is important for the understanding of my findings, is the distinction between *semantics* and *pragmatics*. Now, it is crucial to understand that the semantics-pragmatics distinction is a recurrent debate within the field of linguistics. This debate seems far from being settled. Where should the conceptual line between semantics and pragmatics be drawn? Is it even possible to draw a line between them? If yes, to what extent? How does pragmatics incrementally contribute to semantics? In my thesis, I advocate the relevance-theoretic perspective on this discussion, arguing that the distinction between semantics and pragmatics matches the distinction between linguistically encoded meaning and pragmatically constructed meaning (Hall, 2013). My findings support this notion because they show that the effects of semantic content can be easily altered by means of pragmatic enrichment. This observation is also consistent with the reasoning of Carston (1999), who argues that semantics deals with the construction of representations of arguments and counterarguments, and hence serves inference preparation; pragmatics involves

the manipulation of representations of arguments and counterarguments, and hence serves inference execution. This suggests that the study of rational argumentation benefits from a spotlight on pragmatics, which yields insights above and beyond the limits of a mere semantic interpretation.

The current findings also imply to deepen our thinking about the concepts *argument* and *argumentation*. I deem it fruitful to conceptualize these constructs multidimensionally and diversify potential taxonomies for structuring and categorizing different types of arguments and argumentation. Hornikx and Hahn (2012) state that the study of rational argumentation is diverse in its goals (e.g., argument analysis, argument production, argument evaluation), methods (e.g., analytical, empirical, computational), and disciplinary backgrounds (e.g., philosophy, linguistics, psychology, computer science). Consequently, the taxonomic architectures that constitute the building blocks of theorizing about rational argumentation should reflect this diversity, too. First and foremost, the concept of argumentation in and of itself can be conceived within a meta-conceptual structure of the utilization of reason. Reason can be used retrospectively and prospectively. Retrospective reason includes explanation and justification. Prospective reason includes inquiry and argumentation. This suggests that argumentation may be studied within the context of an overarching meta-concept, namely reason. This conceptualization of argumentation is a bottom-up process that embeds the construct into a wider conceptual matrix. It goes without saying that the conceptualization of argumentation can and should also take place as a top-down process that clarifies the specific instances and subcategories of argumentation. Hahn and Collins (2021) list different possibilities as to how to categorize argumentation. They propose the distinction of *arguments as objects* versus *arguments as process*. Arguments as objects refer to sets of statements, which are typically claims or conclusions that are supported by premises. Arguments as objects

can be divided into deductive arguments (e.g., modus ponens, modus tollens), inductive arguments (e.g., statistical generalization, categorization), abductive arguments (e.g., inference from evidence to causes), and defeasible arguments (e.g., statements being overturned by additional information). Deductive arguments imply that the conclusion must necessarily follow if all premises are true. Inductive arguments, abductive arguments, and defeasible arguments imply that the conclusion is provisional; it could be false even if all premises were true. Arguments as process, by contrast, refer to contextualized arguments that often occur in dialogic settings (e.g., dialectical exchange, integration of proposing and opposing claims, open debate). Although main stream psychology has so far predominantly focused on arguments as objects, arguments as process are increasingly being studied (e.g., Evans et al., 2008; Girotto et al., 2001; Mercier & Sperber, 2011; Sperber et al., 1995; Stevenson & Over, 2001; Walton, 1989, 2008; Walton et al., 2008). My thesis adds to this literature of arguments as process in that it shows when and how people integrate counterarguments by virtue of pragmatic enrichment. In this sense, the insights gained from the current thesis may inform the identification of weak arguments. Weak arguments include: the circular argument, the slippery-slope argument, the analog as argument, the example as argument, the argument from authority, and the ad hominem argument. All these types of arguments are problematic and make for rather weak arguments because they do not focus on the actual content but on irrelevant or at least secondary aspects of an argument. Unlike persuasion, rational argumentation asks what convinces and what *should* convince, thereby combining and reconciling descriptive and normative concerns (Hahn & Oaksford, 2012). A good, rational argument must meet the “burden of proof”. My findings imply that a major player in coming this burden of proof a step closer is relevance.

The knowledge gained from the results of my thesis imply that the concept *relevance* is not only used as a naïve lay construct of everyday folk psychology. Quite the opposite, relevance is extremely powerful as a scientific concept. Its precisely defined and fully spelled out theory as well as the abundance of empirical evidence attest to the concept's validity. In other words—relevance is relevant for rational argumentation. Wilson and Sperber (2012b) argue that relevance offers a pragmatic framework that is of a wider scope than truthfulness, because relevance cannot only account for literal usage of language but can also explain loose and figurative speech acts (e.g., irony, humor, metaphor, politeness, etc.). Consequently, relevance turns out to be a superordinate concept that governs pragmatic aspects of rational argumentation. A crucial auxiliary concept in this regard is *epistemic vigilance*. It helps to explain why humans have the propensity to search for relevant cues of information in the first place. Sperber et al. (2010) articulate the concept as follows: Epistemic vigilance is an evolved, adaptive cognitive mechanism whose function is to protect against misinformation. Arguers often have conflicting interests. When the likelihood of a conflict of interest is high, the own interest is best served by deceiving the interlocutor. Also, deception may not only occur deliberately, but also inadvertently through accidental miscommunication (Sperber, 2013; Wearing, 2015). Therefore, arguers developed the skill of being epistemically vigilant; they monitor the reliability of others' arguments (and counterarguments). Epistemic vigilance, thus, is an important auxiliary concept that complements relevance theory.

Lastly, I wish to briefly sketch a meta-conceptual consideration that is implied by the findings of this thesis. My hope is that the reader of my thesis recognizes that it can be read as a plea for critical rationalism. For the study of rational argumentation, engaging in a mass production of bad research articles is not how the field will thrive. A positivist research program, which solely revolves around empiricist goals, is myopic,

impatient, narrow-minded, and does more harm than good in the long run. As Hilary Putnam (1981) rightfully wrote, it will “produce philosophies which leave no room for a rational activity of philosophy” (p. 113). My thesis constitutes a heartfelt call to revive a psychological science that dignifies its philosophical roots and remembers its historical precursors in order to launch a new renaissance of a theory-driven and critical-rational psychological science. It is my unshakable belief that this is the better way forward!

4.2.3 Methodological Implications

This chapter summarizes the methodological implications of my thesis. I will (a) describe how the employed methods meet the most current requirements of methodological rigor and statistical scrutiny, (b) discuss chances and pitfalls of different types of response time analysis, (c) fit various theoretical distribution models to observed response time data, (d) compare different methods for data transformation of response times, and (e) reflect on the meta-methodological implications.

How do the employed methods meet the most current requirements of methodological rigor and statistical scrutiny? In recent years, as is widely known, psychological science has dealt with what is commonly dubbed replication crisis (Open Science Collaboration, 2015). As a response, there have been many calls for better methods, good practice, ethics, and open science initiatives. Consequently, many new gold standards for good methodological conduct were (re-)formulated. These include, but are not restricted to, power calculations, preregistrations, registered reports, new statistics, open data, open materials, open scripts, replication studies, mixed models, meta-analysis, documentation protocols, online repositories, clear divide between confirmatory hypothesis testing and exploratory hypothesis generation, large-scale collaborations, and more. The way in which I conducted and reported the work of the present thesis satisfies many of these novel methodological developments. I attached all materials (i.e., instructions, stimuli, and response formats) to the appendix of this

thesis. I employed back-translation (Brislin, 1970) to ensure proper translations. I ran power calculations (a-priori power analyses, sensitivity analyses). I documented and stored all experimental programs, power calculations, data, and analysis scripts online in an OSF project. I reported how I determined sample size, all data exclusions (if any), all manipulations, and all measures. I requested and obtained written informed consent from all participants before each experiment. I adhered to the ethical standards of the Declaration of Helsinki (World Medical Association, 2013). After each experiment, all participants were debriefed, compensated, thanked, and dismissed. In conjunction with traditional parameters, I used new statistics (e.g., effect sizes, confidence intervals, Bayes factors, Bayesian sequential analyses). I ran a replication study. I ran a second, extended replication study. I used mixed models to account for the random variance in the multilevel structure of the unaggregated data. I used meta-analyses to assess homogeneity across all experiments and to calculate combined effect size estimates. All these measures that I took to guarantee methodological rigor are strongly encouraged by the scientific community, because they increase the credibility of the conclusions we draw from our research (e.g., see Colling & Szűcs, 2021; Goh et al., 2016; Judd et al., 2012, 2017; Maxwell et al., 2015; Shrout & Rodgers, 2018; Westfall et al., 2014; but also see Fiedler & Prager, 2018). Nosek et al. (2022) rightfully state that replicability, robustness, and reproducibility are improving the quality of psychological research. Replicability refers to testing the reliability of a previous finding with new data. I successfully replicated my findings, as evidenced by the second and the third experiment of my thesis. Robustness is ensured by testing the reliability of a previous finding taking the same data but a different analysis strategy. I successfully accomplished this goal, as evidenced by the mixed model analyses as well as the meta-analyses. Reproducibility refers to testing the reliability of a prior finding using the same data and the same analysis strategy. Reproducibility is also enabled, as I am

open to provide the necessary data and analysis scripts to my colleagues if they want to reproduce the results. Therefore, taken together, the methods and results reported in this thesis imply to offer a replicable, robust, and reproducible data pattern, which warrants the credibility of the scientific conclusions I have drawn.

What are the chances and pitfalls of different types of response time analysis? Response times (also called reaction times or latency) offer an important source of information in cognitive psychology. They can complement other behavioral data, for instance error rates. Thereby, they may validate the conclusions drawn from other measures and assist their interpretation. However, there is an ongoing debate on how to preprocess response time data prior to analysis. Some researchers argue that response times should preferably neither be trimmed nor transformed prior to data analysis, but instead the raw data should be analyzed. Their main argument is that preprocessing of response time data sugarcoats the actually observed distribution by ignoring more extreme values on both tails of the distribution, and by artificially standardizing the observed distribution to a theoretically ideal distribution, which is nothing more than an unwarranted distortion of the actually observed distributional shape. Such practices are arguably pernicious to external validity and a direct manifestation of unjustified assumptions and a form of statistical idealism that fails to live up to the reality of the empirical world. In line with this reasoning, Baayen and Milin (2010) opt for minimal a-priori trimming of response time data; however, if preprocessing is applied to response time data, models for trimming and transforming data should be critically evaluated. Likewise, Lo and Andrews (2015) argue that an eschewal of preprocessing response time data is the better choice in most cases; consequently, they advocate to use the raw response times for analysis. Morís Fernández and Vadillo (2020) used real and simulated data to test the effects of various preprocessing steps on the false-positive rate (i.e., significant findings that

would have remained non-significant if the raw response time data had been used for analysis). They found that when several preprocessing steps were used in combination, the false-positive rate can easily rise up to 17%. Given that more degrees of freedom occur down the analysis pipeline, the final false-positive rate might be even higher. Schramm and Rouder (2019) report that neither an inverse nor a logarithmic transformation of response time data benefits test power or type I error control. In some cases, an inverse transformation even leads to lower power. Considering these findings and recommendations, I deliberately refrained from trimming or transforming my response time data. Instead, adhering to the recommendations outlined above, I used the raw data to analyze the response times. However, I am well aware that other authors proposed various methods for trimming and transforming response time data, respectively. The main tenet of preprocessing is to minimize the effects of outliers. Some methods for the management of outliers are: spotting extreme values by means of boxplots (e.g., Tukey, 1977), transformations (e.g., inverse transformation, logarithmic transformation), trimming a certain percentage of the responses (e.g., 5% of the lower and upper bounds), trimming according to standard deviations (e.g., ± 3 *SD*s), trimming at cutoff values (e.g., < 200 ms), using alternative central tendency parameters (e.g., medians), or using data aggregation (e.g., Vincentizing). It is often unclear which method (if any) a researcher should adopt to preprocess his or her response time data. Ratcliff (1993) gave four recommendations on how to handle preprocessing: (1) try a range of cutoffs and make sure that an effect is significant over some range of non-extreme cutoffs; (2) use the inverse transformation (or standard deviation cutoff if participant variability is large) to confirm the cutoff analyses; (3) if the effect is novel, unexpected, or important, replicate it or partially replicate it in another experiment; and (4) most important, choose the method before analyzing the data, do not use several methods and choose only the one that is significant.

Why fit various theoretical distributions to observed response time data? A valid critique of the herein before mentioned trimming techniques and transformations is that researchers very often base their decision of which one to choose depending on which technique is most common or which they are most familiar with. Obviously, this approach is not the most professional one. Unfortunately, however, this is too often scientific reality. Whelan (2008) suggests an alternative approach that may be more effective when it comes to detecting genuine differences in response times between conditions—namely, analyzing the whole response time distribution. Analyzing the whole response time distribution becomes increasingly popular. Its strength is that it helps to explore effects that would otherwise have been missed. Moreover, theoretical distribution models can be fitted to empirical distributions of response times. Thereby, the model fit of different theoretical distributions with the empirical distribution can be assessed and compared. As a consequence, researchers can then choose their inference test for analyzing the response times according to the statistical assumptions of the theoretical distribution model that yields the best model fit. To demonstrate, I fitted several feasible theoretical distribution models to the unaggregated response time data for modus ponens inferences and modus tollens inferences, combined across all experiments, all participants, and all other conditions, resulting in $N = 2400$ observations per inference type. Data and analysis code for the modeling are documented and stored in the thesis' OSF project (https://osf.io/3dm2j). I provide access upon reasonable request. Table 22 shows the descriptive statistics of the empirical response time distributions for modus ponens and modus tollens inferences. Figure 28 shows skewness-kurtosis plots as proposed by Cullen and Frey (1999) to indicate candidate distributions. In order to take into account the uncertainty of the estimated skewness and kurtosis values, I performed a non-parametric bootstrap procedure (Efron & Tibshirani, 1994), which computed skewness and kurtosis values

based on *k* = 1000 bootstrap samples (i.e., random sampling with replacement from the original data set) per inference type. Based on the inspection of the skweness-kurtosis plots, the bootstrapping, and general knowledge on the distributional properties and the governing stochastic processes of response time as a random variable, I chose the three following theoretical distribution models as likely candidates to fit the empirically observed response time distributions: the lognormal distribution, the normal distribution, and the Weibull distribution. Therefore, I fitted these three distribution models to the data. Table 23 shows the model parameters and fit indices of the lognormal model, the normal model, and the Weibull model for the response time distributions of modus ponens inferences and modus tollens inferences. Figure 29 shows the distribution model fits for modus ponens and modus tollens. The lognormal model yields the best fit to the empirical data. The Weibull model yields the second best fit. The normal model yields the worst fit.

Table 22. Descriptive statistics of the response time distributions for modus ponens inferences and modus tollens inferences.

	MP	MT
N	2400	2400
M	3580	4465
Mdn	2700	3349
SD	3096	4622
Min	292	364
Max	39435	103852
Skewness	3.89	7.87
Kurtosis	27.88	122.41

Note. MP = modus ponens; MT = modus tollens.

How to compare different methods for data transformation of response times? There is no consensus on whether and how to transform response time data before analysis. Each strategy has advantages and disadvantages that should be considered.

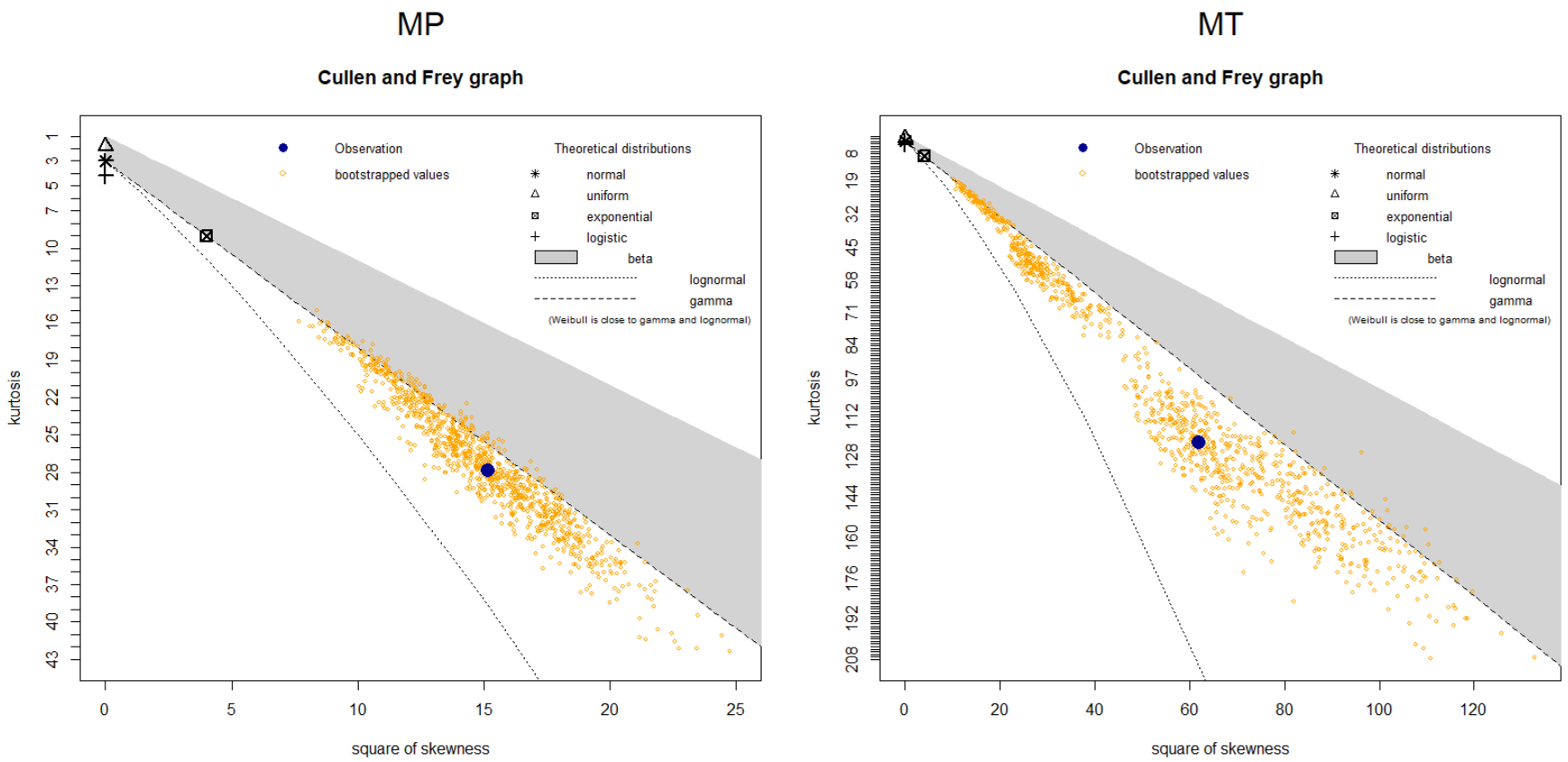

Figure 28. Cullen and Frey graphs, showing kurtosis-skewness relation of the observed distribution (*N* = 2400), bootstrapped values (*k* = 1000), and theoretical distributions. Left panel: modus ponens inferences. Right panel: modus tollens inferences.

Table 23. Model parameters and fit indices of the lognormal model, the normal model, and the Weibull model for the response time distributions of modus ponens inferences and modus tollens inferences.

Model	MP	MT
Lognormal		
Parameters		
log(*M*)	7.95 (0.01)	8.15 (0.01)
log(*SD*)	0.65 (0.01)	0.67 (0.01)
Fit indices		
Log-likelihood	-21467.03	-22004.96
AIC	42938.07	44013.93
BIC	42949.63	44025.50
Normal		
Parameters		
M	3579.96 (63.12)	4464.67 (94.16)
SD	3095.44 (44.67)	4620.88 (66.58)
Fit indices		
Log-likelihood	-22695.90	-23657.47
AIC	45395.79	47318.94
BIC	45407.36	47330.51
Weibull		
Parameters		
a	1.39 (0.02)	1.30 (0.02)
b	3972.11 (61.99)	4893.52 (82.12)
Fit indices		
Log-likelihood	-21810.63	-22412.02
AIC	43625.26	44828.03
BIC	43636.83	44839.6

Note. Standard errors are in parentheses. The distribution models were fitted using maximum likelihood estimation (MLE). *N* = 2400 observations. MP = modus ponens; MT = modus tollens. *a* = shape parameter; *b* = scale parameter.

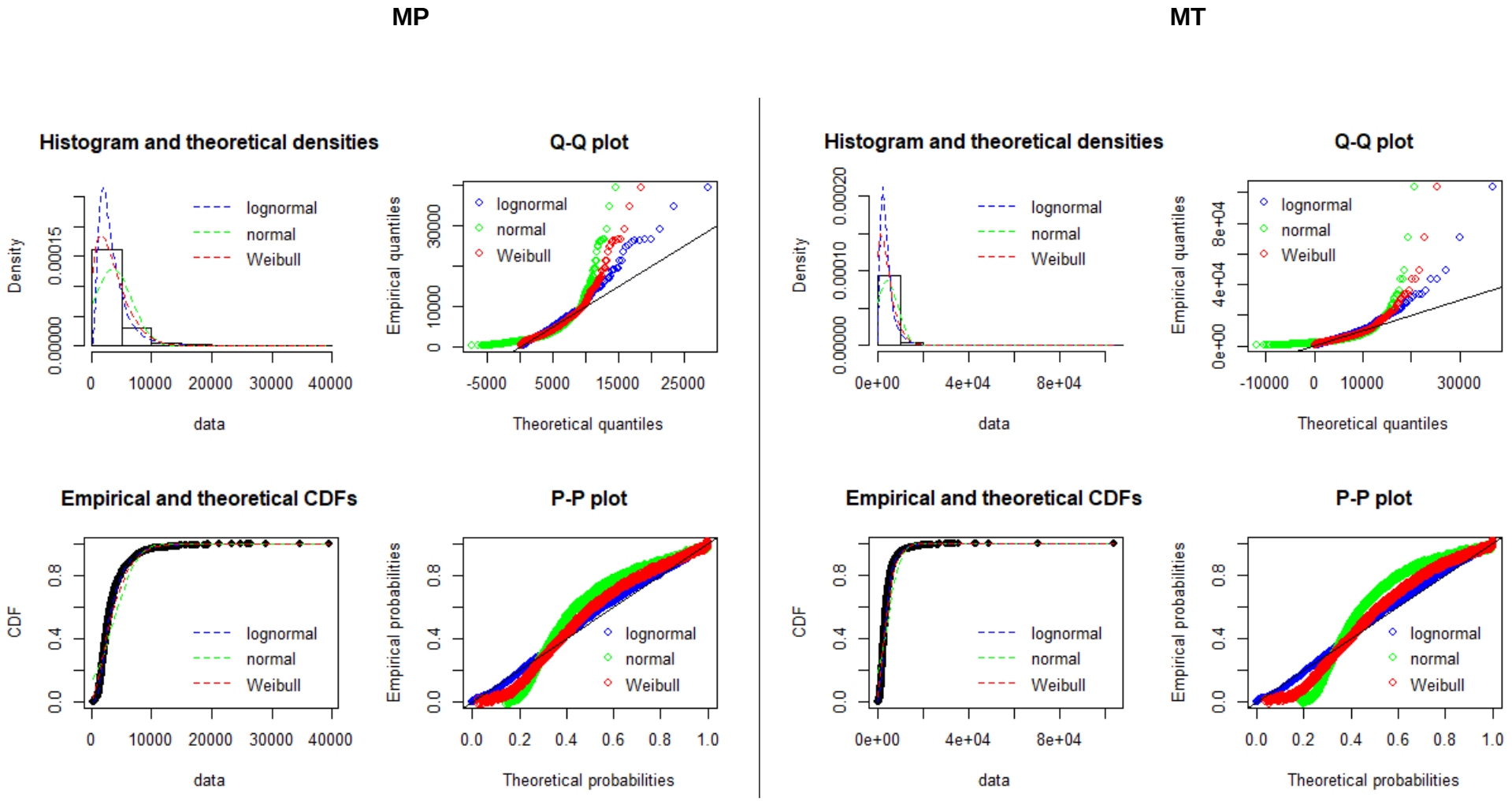

Figure 29. Distribution model fits, visualized by density plot, Q-Q plot, CDF plot, and P-P plot. Left panel: modus ponens inferences. Right panel: modus tollens inferences. The distribution models were fitted using maximum likelihood estimation (MLE).

I left the response time data untransformed. The more preprocessing steps one applies to raw data, the less likely it is for the processed data to provide an accurate empirical description of the data that actually have been observed. Response time transformations should not be done for the sake of creating a data set that can be easily analyzed with the most common statistic. Data should not be fitted to statistical models. Instead, statistical models should be fitted to data. Therefore, in most cases it is arguably the better strategy to choose a model whose assumptions do not violate the distributional characteristics of the data, even if this means that the data must be analyzed with non-parametric inference tests. Moreover, if the response time data display a right-skewed distribution, which is typical for response times, then why should this distribution artificially be distorted? Since this was the case for the response times reported in this thesis, I did not transform them. However, in other cases transformation can be justified, for instance when distributions have been deformed blatantly by stark outliers. Common response time transformations are the inverse-transformation (i.e., 1/RT) and the log-transformation (i.e., log[RT]). Figure 30 shows the impact of these transformations on my response time data. I used the unaggregated response time data for modus ponens inferences and modus tollens inferences, combined across all experiments, all participants, and all other conditions, resulting in N = 4800 observations in total. Data and analysis code for the transformations are documented and stored in the thesis' OSF project (https://osf.io/3dm2j), to which I provide access upon reasonable request. As can be seen, the untransformed response times (left panel) show the typical right-skewed distribution, both for modus ponens inferences as well as modus tollens inferences. Variability is smaller for modus ponens response times than for modus tollens response times. The inverse-transformed response times (middle panel) remain also right-skewed; however, variability of modus ponens and modus tollens response times is reversed. The log-transformed response times (right

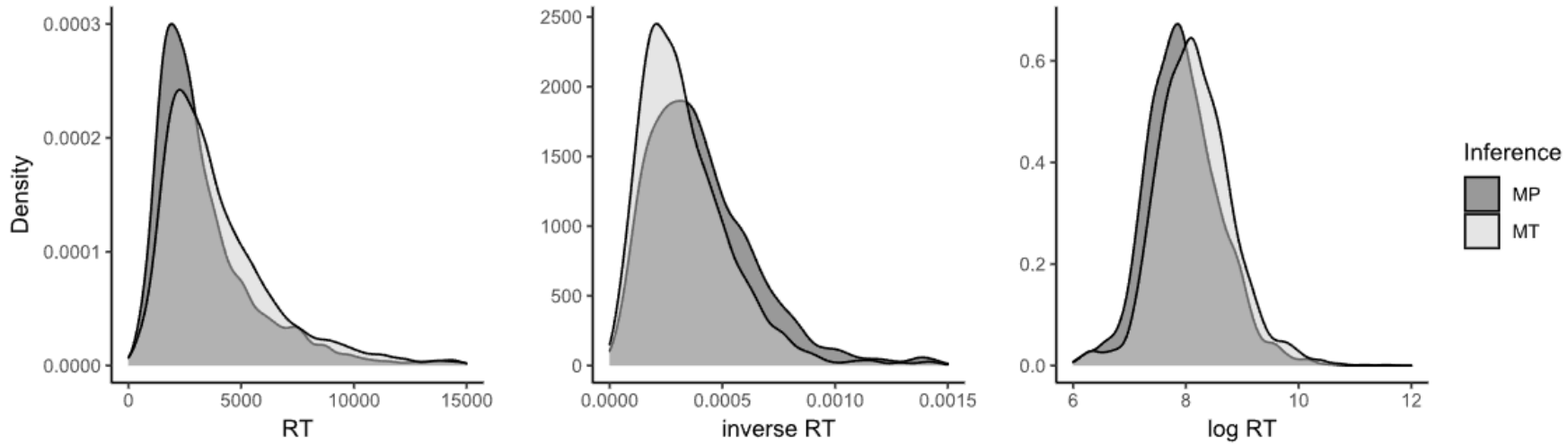

Figure 30. Density plots for the response time distributions of modus ponens inferences and modus tollens inferences. Left panel: untransformed response times (ms). Middle panel: inverse-transformed response times. Right panel: log-transformed response times. RT = response time; MP = modus ponens; MT = modus tollens. *N* = 4800 observations.

panel) approximate normal distributions; also, variabilities of modus ponens and modus tollens response times are virtually equal. This example demonstrates how transformations can change response time distributions. Obviously, this invalidates the conclusions one draws from the analysis of response time data. Therefore, response times should only be transformed when one has good reasons to do so. If such reasons are absent and the raw data pattern follows the pattern that is theoretically expected for response times, namely a right-skewed distribution, then, in most cases, it is the better strategy to leave the data untransformed.

What are the meta-methodological implications? An essential meta-methodological implication of my thesis is that a pluralistic approach to research methodology enriches psychological science and is superior to a rigid and narrow-minded focus on a single paradigm. The diversification of experimental paradigms, statistical testing, and modeling of cognitive processes is a direct manifestation of scientific freedom. And scientific freedom also includes freedom of methods! Applying a wide range of cognitive-psychological methods and developing new paradigms is challenging, but it definitely promises rich rewards for the study of rational argumentation (Hahn & Collins, 2021). Another implication is that taking the time to think about the philosophical foundations of a specific research program will ultimately help to improve its methods, too. This reasoning is in line with the postulations of the emerging field of meta-science, which combines insights from philosophy of science and theory of science with the study of scientific methods, in order to create new knowledge that helps to improve the quality of science as a whole itself (Schooler, 2014). My findings also show that a clearly spelled-out theory and a valid operationalization via a good experimental design make collecting data from thousands of participants a waste of time and money. If theory and design align tightly, then small samples are sufficient. Increasing test power simply by increasing sample

size does not impress me much. In fact, if an effect can only be demonstrated with a huge sample (but not with a smaller sample or with several smaller samples), it is rather an alarm signal that the effect itself is a dirt effect that is of no practical significance whatsoever. On a related note, it should also be mentioned that power analysis for sample size planning should always be conducted with a critical mind. A power analysis is nothing more than a statistical heuristic—it is not deterministic! Consequently, I argue that a power analysis may be used for sample size planning, but it should never be the only criterion to determine sample size (for a similar argument, see Fiedler, 2020a). Instead of overestimating the significance of power analysis, a meta-methodological aspect that deserves more attention refers to manipulation checks. Manipulation checks are important because they indicate whether or not an experimental manipulation was successful and hence whether construct validity is given (Fiedler et al., 2021). A final meta-methodological reflection addresses the mere reliance of psychological science on experiments. If one takes the plea for a pluralistic approach to scientific psychology seriously, then other means of creating knowledge should be integrated more into psychology—not just in principle, but also in practice. These may include, but are not restricted to, macro-scale surveys, micro-scale observations, ethnographic field work, archival inspection, interviews, correlational studies, philosophical inquiry, literature analysis, longitudinal studies, computational modeling, neural networks, and sociological network analysis. For a recent article that expresses the idea of a multi-methodological approach, I refer the interested reader to Diener et al. (2022).

4.2.4 Practical Implications

This chapter summarizes the ways in which the insights gained from this thesis can help to inform practical applications of rational argumentation for interventions, remedies, training programs, and other forms of change programs in various societal

structures and institutional settings. The body of research on the applied value of rational argumentation is large. Therefore, this chapter will not treat all, but the most important and influential avenues that have been studied.

First, rational argumentation has a tremendous value for building applications in the field of computer science, especially within artificial intelligence (see Bench-Capon & Dunne, 2007). Implementing the rules of rational argumentation into computer programs is supposed to guide decision processes in real-world scenarios. For example, aspects of rational argumentation are being considered when building legal expert systems, which are algorithms aimed to give legal advice (e.g., Prakken, 2008). Argumentation-based frameworks are also increasingly implemented in computational medical decision support systems (e.g., Fox et al., 2007) as well as multi-agent systems (e.g., Rahwan & Moraitis, 2009). There are also software packages that were developed to visualize complex sequences of arguments, like the argument mapping software Araucaria (Reed & Rowe, 2004). Other software programs allow web users to analyze arguments on a particular topic across large-scale distributed online content (e.g., Rahwan et al., 2007). Hahn and Oaksford (2012) argue that the practical value of all these computational implementations lies in the fact that they assist argumentation-based decision processes by capturing the dialectical structure and relationships of theses (i.e., arguments), antitheses (i.e., counterarguments), and synthesis (i.e., conclusion).

Second, rational argumentation can be trained and improved during group discussions and open debates. It has been demonstrated that regular discursive practice sharpens argumentation skills and even increases overall reasoning performance (e.g., Bonner et al., 2002; Laughlin & Ellis, 1986; Stasson et al., 1991). Rational argumentation also fosters cooperation because reasoning in argumentative contexts channels the benefits of logical and systematic thinking to serve an ultimately

social function, namely producing strong arguments yourself and evaluating the strength of others' arguments. Thereby, rational argumentation is an excellent exercise both for reasoning and social competence (Sperber & Mercier, 2012). When people engage in open debates, what usually happens is that the validity of their arguments is put to the test by others and vice versa. Consequently, the weak arguments will be withdrawn, whereas strong arguments will survive the argumentative confrontation. In the end, the best argument will push through. Intervention studies suggest that argumentative intervention programs increase the frequency of attempts to make arguments, justify them, respond to other arguments, and compare competing arguments (Crowell & Kuhn, 2014; Hahn & Collins, 2021). Intervention program for training rational argumentation can be adjusted towards specific time frames and participant populations. Those interventions may range from an hourlong program for undergraduates (Zavala & Kuhn, 2017) to a three-year program with children (Crowell & Kuhn, 2014; Kuhn & Crowell, 2011). It strikes me as fascinating to learn that the training of rational argumentation displays transfer effects to other domains of intellectual challenge. For instance, Nussbaum and Asterhan (2016) have shown that a training in mathematical argumentation led to higher test scores on standardized tests of mathematics and reading. This finding suggests that rational argumentation functions as a powerful, highly generalized intellectual tool that is beneficial for a wide variety of cognitive domains in which humans excel. The elicitation of transfer effects of rational argumentation by means of intervention programs does not only improve immediate reasoning performance; it also improves long-term reasoning performance (Resnick et al., 2013).

Third, the insights gained from the present thesis complement knowledge on how rational argumentation can be made useful in education. This includes early childhood development, formal schooling, higher education, and professional science

training. Research into rational argumentation in children can help to understand how the argumentative mind is shaped during the early stages of human development and illuminates how to foster argumentation skills in young children (Anderson et al., 1997; Brem & Rips, 2000; Genishi & DiPaolo, 1982; Glassner et al., 2005; Kuhn, 1989, 1991, 2001; Means & Voss, 1996). This, in turn, will be fruitful for devising playful learning approaches for young children that are fun, engaging, and motivate children to keep being curious learners throughout their lives. In particular, training children in rational argumentation is especially beneficial in science education (Kuhn, 1993). It has been demonstrated that the nurturance of rational argumentation in school children improves their scientific literacy, for example, their understanding and usage of hypotheses and evidence (Klaczynski, 2000; Kuhn et al., 2000; Kuhn & Udell, 2003; Norris & Phillips, 2003; Sadler, 2004). The study of rational argumentation does not only have practical implications for children and adolescents, but also for advancing our knowledge about how people learn in higher education. Specifically, insights in aspects of rational argumentation are practically used in philosophy of science in order to better understand how exactly scientists construct hypotheses, design experiments, and analyze data (Bovens & Hartmann, 2003; Earman, 1992; Howson & Urbach, 2006).

Fourth, rational argumentation has practical implications for law. It is an invaluable learning technique applied in moot court competitions among law students. The students receive a case, study it, and then simulate a legal trial. One group of students is assigned to the side of the defense lawyer; another group of students is assigned to the side of the prosecutor. Each group must produce arguments for their side. The other group has to evaluate those arguments and produce counterarguments, and so on and so forth. Then, the groups switch their roles. This technique turns out to be highly effective in training law students in argumentative reasoning—a skill arguably being the number one predictor of success in their later

professional lives. After all, a lawyer's job as advocate is not to lead the judge of a court trial to some "absolute truth", but to convince the judge that his or her client's position is *more* legally and logically correct than the opposing side's position (O'Neill, 2012). With respect to judges, Kahan (2011) argues that they should cultivate the quality of *aporia* to guide their decisions, which refers to the awareness that different perspectives on a given case inevitably produce different outcomes. A judge must always keep this in mind in order to stay objective, independent, and neutral. Research with mock juries has shown that jury decision making is strongly influenced by the way that prosecution and defense build up their argumentative narratives of the case, that is, how coherently they describe and explain the sequence of initiating events, goals, actions, consequences, and accompanying events of a given case (Pennington & Hastie, 1981). The more logically structured these argumentative stories are, the more memorable and impactful they are for the jury members' decision (Pennington & Hastie, 1981, 1986, 1988, 1992; Voss & Van Dyke, 2001). Another factor that co-determines jury decisions is the comprehensiveness of the opening statements of prosecution and defense, respectively, because it strongly influences the processing of the subsequent arguments expressed by both parties (Pyszczynski & Wrightsman, 1981; see also Pyszczynski et al., 1981).

Fifth, rational argumentation has practical implications for policy making. Liberal democracy crucially relies on the free expression of dissenting opinions (Mercier & Landemore, 2012). Free expression of dissent constitutes a prerequisite for political decision processes like making new laws in an open society. It fulfills the key function of democracy, namely to distribute power so that no single authority has to much power in its hands. Power must be limited—and this is achieved best by strong counterarguments. Furthermore, a central property of good parliamentary practice consists in deliberation, both within and across political parties. Rational argumentation

fosters the epistemic standards of such deliberations by increasing respect between interlocuters (Gutmann & Thompson, 1996; Schneiderhan & Khan, 2008; Steenbergen et al., 2003), increasing the likelihood to ultimately reach consensus (Dryzek & Niemeyer, 2006; Niemeyer & Dryzek, 2007), and increasing coherence between beliefs (Gastil & Dillard, 1999). Liberal democracy is characterized by a discursive, dialogical ideal of deliberation, that is, talking out conflicting views with others (Landemore & Mercier, 2012). Another positive effect of rational argumentation in deliberative contexts is that it helps to establish and maintain symmetrical relationships between all interlocuters (Habermas, 1981, 1992). It has been shown that heterogeneous groups make for better rational-argumentative deliberation, because groupthink is less likely to occur and the free exchange of opposing views and alternative perspectives is more frequent (e.g., Caluwaerts & Deschouwer, 2014; Suiter et al., 2021). Forgas and Lantos (2020) ascertain that "[…] Plato noted more than 2000 years ago, one of the greatest dangers for democracy is that ordinary people are too easily swayed by the emotional and deceptive rhetoric of ambitious politicians" (p. 287). Unfortunately, people are evolutionarily predisposed to attend to emotionally appealing narratives that enhance ingroup favoritism and outgroup derogation (Gelfand & Lorente, 2021; Harari, 2014). However, humans also have the capacity for rational argumentation when they are willing to invest some effort. It is precisely through this deliberate investment that rational argumentation helps to reaffirm the humanist Enlightenment values of autonomous individualism and liberal democracy.

Finally, in recent years, we must ascertain that academic freedom is increasingly being threatened by an authoritarian and intolerant ideology that claims a cultural hegemony in public life. Unfortunately, this ideology is massively reinforced by biased media reports and regulatory overkill from governmental institutions. Proponents of this movement often use infamous arguments to morally condemn their

interlocuters. This development has to stop! If we wish to create a healthy intellectual culture that values diverse perspectives and tolerates deviant opinions, we must ignore *arguments from ideology* but reward *arguments from reason*. Rational argumentation will not only improve the quality of academic scholarship, it will also strengthen open societies. After all, it is rational argumentation that enables peaceful negotiations of conflicting interests between individuals, groups, organizations, and states. Thereby, rational argumentation consolidates the values of open societies and counteracts totalitarian tendencies (Popper, 1945).

4.3 Limitations and Future Directions

In this chapter, I will name some limitations and boundaries of the present work. I will connect this reflection with an outlook on potential avenues for future research on rational argumentation.

The experiments reported in this thesis were conducted in Western societies, testing Central and South European participants. It is questionable whether the results would generalize over other populations. My hypothesis is that the detected pattern of results would replicate amongst North American and Australian study participants, and also in other European countries. The populations inhabiting the listed world regions are mostly Western, educated, industrialized, rich, and democratic societies, which cherish the values of open societies and liberal democracy. These values comprise an appreciation for a dialectical resolution of conflicting interests via the exchange of rational arguments in open debates. Therefore, I assume that the processes by which members of these populations integrate arguments and counterarguments to reach a conclusion should work rather similar. However, it would be an interesting task for future research to test whether or not the observed pattern remains robust in Eastern samples (e.g., East Asia, South Asia) and in samples that are often neglected altogether in psychological research (e.g., Latin America, Africa). In those participant

pools, I am less certain whether to expect replication or not. From a general psychological standpoint, one could argue that rational argumentation is a manifestation of rational thought, which is an inherently universal feature of the human mind. From this point of view, one would predict a generalization of the findings. However, from a social psychological standpoint, one could argue, for instance, that Eastern societies value politeness, social stability, harmonious relationships, and the conservation of social hierarchies. Consequently, study participants from Eastern cultures might regard, process, and integrate counterarguments differently, or at least to a different extent. Therefore, future research could take a closer look at when and how cultural variables affect rational argumentation.

A few authors (Darmstadter, 2013; Dogramaci, 2020; Pérez Zafrilla, 2016; Sterrett, 2012) raised concerns with respect to the scope of the argumentative theory of reasoning. For example, one critique is that the theory does not adequately distinguish between normative accounts of rational argumentation (i.e., how one *should* reason in argumentative contexts) and descriptive accounts of rational argumentation (i.e., how one reasons in argumentative contexts). I consider this a valid criticism that should be taken into account thoroughly in order to advance the theory. For instance, with regards to argument evaluation, an interesting future task is to exactly define the normative standards for the rational evaluation of arguments. This is important because it provides us with formal criteria against which to measure rational argumentation during argument evaluation. As a consequence, these normative standards will not only sharpen the theory; they will also improve the measurement of rational argumentation. On the implementational level of measurement, Prado et al. (2020) found that the medial prefrontal cortex is connected to argumentative reasoning. However, the neural basis of rational argumentation, as of yet, is far from being completely understood.

A boundary condition of the present work is that it specifically focused on rational argumentation during argument evaluation. A potential extension of this project could be to dig into rational argumentation during argument production. The argumentative theory of reasoning would predict a strong confirmation bias during argument production. One interesting task would be to study interindividual differences with respect to the probability of occurrence of a confirmation bias during argument production. Interestingly, and somewhat surprisingly, the confirmation bias has been shown to be virtually unrelated to intelligence (Stanovich et al., 2013). This suggests that interindividual variability of the confirmation bias during argument production should be small. However, rationality is a larger construct than intelligence; it entails aspects that intelligence tests miss to capture (Stanovich, 2009, 2011). Future research might attempt to develop measurement tools that capture aspects of rationality that conventional IQ tests miss. It would be interesting to see whether variance in test scores on such measures were correlated with different degrees to which people display a confirmation bias during argument production. Moreover, confirmation bias may be more or less prevalent in argument production depending on the specific cognitive processing stage. Vedejová and Čavojová (2022) found that confirmation bias was present during information search, whereas it was less prevalent during interpretation, and even absent during memory recall. This suggests that confirmation bias arguably exerts differential effects on argument production, depending on the specific processing stage at work. Furthermore, future research can continue to collect evidence showing that the confirmation bias is actually an adaptive feature of argument production rather than a problem. For example, Rollwage and Fleming (2021) used simulation-based modeling to demonstrate how the confirmation bias can serve as an adaptive tool given it is coupled with good metacognitive strategies. Being equipped with good metacognition allows reasoners to utilize

confirmation bias in order to attach a lower weight to contradictory information when they are correct, but still seek new information when they realize they are wrong.

Another idea for future research concerns the connection between rational argumentation and working memory. Shehab and Nussbaum (2015) compared the cognitive load involved when using different critical thinking strategies during argument-counterargument integration (Nussbaum, 2008), which describes the process by which reasoners evaluate, refute, and synthesize two sides of an issue in order to reach a justification for an overall conclusion. The two critical thinking strategies they compared were (a) constructing design claims that minimize the disadvantages of an alternative, and (b) weighing refutations which weaken an argument by arguing that there are more important values at stake. Weighing refutations was connected to more mental effort than constructing design claims. This relationship was especially pronounced for participants who scored high on the Need for Cognition (NFC) scale (Cacioppo et al., 1984). Presumably, weighing refutations demands more cognitive load because disparate elements have to be coordinated in parallel in working memory, whereas constructing design claims is a sequential process that needs less cognitive scaffolding. Oberauer and Greve (2021) showed in a series of eight experiments that maintenance of information in working memory is highly selective. It is controlled by current goals and what is considered relevant to achieve these goals. Further elucidating the cognitive mechanisms and capacity constraints by which relevant arguments and counterarguments are processed in working memory is an extremely interesting and engaging task for future research.

Prospectively, conducting research on rational argumentation through the lens of joint action seems promising for being yet another productive research enterprise. Joint action describes “any form of interaction involving at least two agents that is made fully intelligible by reference to representational features accessed by the subject in

the first-person plural" (Gallotti & Frith, 2013, p. 160). Developing new experimental designs to study rational argumentation in this *we-mode* would be an interesting and challenging task for the future. Joint action research has already generated many intriguing and highly creative paradigms (Sebanz & Knoblich, 2021). These paradigms could be innovated so as to tailor them to the study of joint rational argumentation. Herbert Simon's seminal work on the two models of man—man as a social being, and man as a rational being—have long constituted a duality (Simon, 1957). Combining rationality research and joint action research may integrate and reconcile these models into a unified model of social rationality (Lindenberg, 2001).

Future research on rational argumentation might also benefit from adopting a Bunswikian approach of representative design (Brunswik, 1955, 1956). Representative design takes into account the notion that cognitive functioning is adapted to the structural features of the environment in which it is embedded (Dhami et al., 2004; Fiedler, 2020b). A central advantage of representative design is that it acknowledges the relations between different dimensions inherent in a psychological construct and deliberately intends to preserve them within the experimental design. This allows to investigate how the different layers of a multidimensional construct co-determine the measured outcome in externally valid contexts. I have already demonstrated how a Brunswikian sampling approach can successfully be utilized for creating a representative design in order to study the multidimensional learning of affective meaning (Richter & Hütter, 2021). I have also contemplated on how representative design can inspire innovations in attitude research (Richter, 2021). Likewise, I do believe that the study of rational argumentation can benefit from the implementation of representative designs. For instance, we found a negativity bias in conditional reasoning with counterarguments in an orthogonal, fully crossed factorial design (Gazzo Castañeda et al., 2016). It would be an interesting task for future research to

test whether this negativity bias remains robust, whether it vanishes, or whether it is enhanced under the circumstances of a representative design.

Finally, on a general note, I think that future research should revisit the reasons why psychology suffers from a replication crisis. I am well aware that it is trendy to tackle psychology's replication problems by means of a stiff focus on statistical modeling and transparency practices. However, I think that this approach, if anything, only cures the symptoms. It does not go down to the root of the problem. The true problem lies elsewhere: Researchers do not take the time to logically derive a testable hypothesis from a theory. If researchers tested hypotheses that imply a strong link to the underlying theory, then the rate of Type I errors would dramatically reduce. Why? Because such hypotheses are falsifiable. I highly recommend the ingenious work by Oberauer and Lewandowsky (2019) on this issue.

4.4 Conclusion

The research objective of the present thesis was to study the pragmatic modulation of rational argumentation in conditional reasoning with counterarguments. Based on a theoretical integration of relevance theory and the argumentative theory of reasoning, I accumulated evidence for the impact of inference type and linguistic mode on rational argumentation. Importantly, I found confirmatory evidence that the inference type of the conditional and the linguistic mode of the counterargument jointly predict conclusion endorsement. Specifically, conclusion endorsement for modus ponens inferences is higher than for modus tollens inferences when subjunctive counterarguments are present. In contrast, conclusion endorsement for modus ponens inferences is lower than for modus tollens inferences when indicative counterarguments are present. Across three experiments, a reanalysis using mixed models, and an integrative data analysis using meta-analyses, I have shown that this pattern is replicable, reproducible, and robust. Moreover, the pattern reemerged in

different language-culture groups, which provides tentative support for its invariance across languages. I further identified relevance as a crucial boundary condition for the observed interaction of inference type and linguistic mode. In addition, I measured response times to function as a marker for the underlying cognitive mechanisms. While the findings of this thesis provide a clear picture of some aspects of rational argumentation on the functional level, future research must continue to elucidate the cognitive mechanisms on the algorithmic level. Generally, the findings suggest that the exact framing of arguments and counterarguments influences the conclusiveness of an argumentative inference. Rational argumentation is shaped both by logical norms and pragmatic principles. This suggests that an integrated normative-descriptive model is the best way forward to achieve an erudite and profound understanding of rational argumentation. After all, the human mind is not only an information processor. It is a creator of meaning (Bruner, 1990).

References

Adams, E. W. (1975). *The logic of conditionals.* Reidel.

Adams, E. W., & Levine, H. P. (1975). On the uncertainties transmitted from premises to conclusions in deductive inferences. *Synthese, 30*(3), 429–460.

Alavi, M., Archibald, M., McMaster, R., Lopez, V., & Cleary, M. (2018). Aligning theory and methodology in mixed methods research: Before design theoretical placement. *International Journal of Social Research Methodology, 21*(5), 527–540.

Allott, N. (2013). Relevance theory. In A. Capone, F. Lo Piparo, & M. Carapezza (Eds.), *Perspectives on linguistic pragmatics* (pp. 57–98). Springer.

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician, 73*(S1), 262–270.

Anderson, R. C., Chinn, C., Chang, J., Waggoner, M., & Yi, H. (1997). On the logical integrity of children's arguments*. Cognition and Instruction, 15*(2), 135–167.

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research, 52*(3), 305–324.

Atlas, J. D. (2005). *Logic, meaning, and conversation: Semantical underdeterminacy, implicature, and their interface.* Oxford University Press.

Austin, J. L. (1962). *How to do things with words.* Oxford University Press.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28.

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*(1), 206–234.

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*(3), 666–678.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). *Parsimonious mixed models.* arXiv. https://doi.org/10.48550/arXiv.1506.04967

Baum, L. A., Danovitch, J. H., & Keil, F. C. (2008). Children's sensitivity to circular explanations. *Journal of Experimental Child Psychology, 100*(2), 146–155.

Baumeister, R. F., & Masicampo, E. J. (2010). Conscious thought is for facilitating social and cultural interactions: How mental simulations serve the animal–culture interface. *Psychological Review, 117*(3), 945–971.

Behr, D. (2017). Assessing the use of back translation: The shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology, 20*(6), 573–584.

Beller, S., & Spada, H. (2003). The logic of content effects in propositional reasoning: The case of conditional reasoning with a point of view. *Thinking and Reasoning, 9*(4), 335–378.

Belligh, T., & Willems, K. (2021). What's in a code? The code-inference distinction in neo-Gricean pragmatics, relevance theory, and integral linguistics. *Language Sciences, 83*(1), Article 101310.

Bench-Capon, T. J., & Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial Intelligence, 171*(10), 619–641.

Bennett, J. (2003). *A philosophical guide to conditionals.* Oxford University Press.

Berto, F., & Özgün, A. (2021). Indicative conditionals: Probabilities and relevance. *Philosophical Studies, 178*(11), 3697–3730.

Billig, M. (1996). *Arguing and thinking: A rhetorical approach to social psychology.* Cambridge University Press.

Blakemore, D. (2001). Discourse and relevance theory. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 100–118). Blackwell.

Blanchette, I. (2006). Snakes, spiders, guns, and syringes: How specific are evolutionary constraints on the detection of threatening stimuli? *Quarterly Journal of Experimental Psychology, 59*(8), 1484–1504.

Blanchette, I., & Campbell, M. (2012). Reasoning about highly emotional topics: Syllogistic reasoning in a group of war veterans. *Journal of Cognitive Psychology, 24*(2), 157–164.

Blanchette, I., & Caparos, S. (2013). When emotions improve reasoning: The possible roles of relevance and utility. *Thinking and Reasoning, 19*(3), 399–413.

Blanchette, I., Gavigan, S., & Johnston, K. (2014). Does emotion help or hinder reasoning? The moderating role of relevance. *Journal of Experimental Psychology: General, 143*(3), 1049–1064.

Blanchette, I., & Leese, J. (2011). The effect of negative emotion on deductive reasoning. *Experimental Psychology, 58*(3), 235–246.

Blanchette, I., & Richards, A. (2004). Reasoning about emotional and neutral materials: Is logic affected by emotion? *Psychological Science, 15*(11), 745–752.

Blanchette, I., & Richards, A. (2010). The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning. *Cognition and Emotion, 24*(4), 561–595.

Blanchette, I., Richards, A., Melnyk, L., & Lavda, A. (2007). Reasoning about emotional contents following shocking terrorist attacks: A tale of three cities. *Journal of Experimental Psychology: Applied, 13*(1), 47–56.

Bonnefon, J. F., & Hilton, D. J. (2004). Consequential conditionals: Invited and suppressed inferences from valued outcomes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 28–37.

Bonnefon, J. F., & Vautier, S. (2010). Modern psychometrics for the experimental psychology of reasoning. *Acta Psychologica Sinica, 42*(1), 99–110.

Bonnefon, J. F., & Villejoubert, G. (2007). Modus tollens, modus shmollens: Contrapositive reasoning and the pragmatics of negation. *Thinking and Reasoning, 13*(2), 207–222.

Bonner, B. L., Baumann, M. R., & Dalal, R. S. (2002). The effects of member expertise on group decision-making and performance. *Organizational Behavior and Human Decision Processes, 88*(2), 719–736.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97–111.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). Wiley.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the

analysis of a single neuroimaging dataset by many teams. *Nature, 582*(7810), 84–88.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology.* Oxford University Press.

Bowerman, M., & Levinson, S. C. (Eds.). (2001). *Language acquisition and conceptual development.* Cambridge University Press.

Braine, M. D., & O'Brien, D. P. (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98*(2), 182–203.

Braine, M. D., & O'Brien, D. P. (Eds.). (1998). *Mental logic.* Erlbaum.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van‘t Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*(1), 217–224.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*(3), 333–342.

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive Science, 24*(4), 573–604.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185–216.

Brislin, R. W. (1976). Comparative research methodology: Cross-cultural studies. *International Journal of Psychology, 11*(3), 215–229.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. L. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–164). SAGE Publications.

Bruner, J. S. (1990). *Acts of meaning.* Harvard University Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*(3), 193–217.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* University of California Press.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*(1), 61–83.

Byrne, R. M. J. (1991). Can valid inferences be suppressed? *Cognition, 39*(1), 71–78.

Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences, 6*(10), 426–431.

Byrne, R. M. J. (2005) *The rational imagination: How people create alternatives to reality.* The MIT Press.

Byrne, R. M. J. (2007) Précis of The rational imagination: How people create alternatives to reality. *Behavioral and Brain Sciences, 30*(1), 439–480.

Byrne, R. M. J., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language, 40*(3), 347–373.

Byrne, R. M. J., & Johnson-Laird, P. N. (2009). 'If' and the problems of conditional reasoning. *Trends in Cognitive Sciences, 13*(7), 282–287.

Byrne, R. M. J., & Walsh, C. R. (2002, August 7–10). *Contradictions and counterfactuals: Generating belief revisions in conditional inference* [Paper presentation]. 24th Annual Conference of the Cognitive Science Society, Fairfax, VA, USA.

Cacioppo, J. T., & Petty, R. E. (1979). Effects of message repetition and position on cognitive response, recall, and persuasion. *Journal of Personality and Social Psychology, 37*(1), 97–109.

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306–307.

Caluwaerts, D., & Deschouwer, K. (2014). Building bridges across political divides: Experiments on deliberative democracy in deeply divided Belgium. *European Political Science Review, 6*(3), 427–450.

Carroll, J. (1966). Quelques measures subjectives en psychologie: Fréquence des mots, significativité et qualité de traduction. *Bulletin de Psychologie, 19*(1), 580–592.

Carston, R. (1999). The semantics/pragmatics distinction: A view from relevance theory. In K. Turner (Ed.), *The semantics/pragmatics interface from different points of view* (pp. 85–125). Elsevier.

Castelain, T., Bernard, S., Van der Henst, J. B., & Mercier, H. (2016). The influence of power and reason on young Maya children's endorsement of testimony. *Developmental Science, 19*(6), 957–966.

Chan, D., & Chua, F. (1994). Suppression of valid inferences: Syntactic views, mental models, and relative salience. *Cognition, 53*(3), 217–238.

Channon, S., & Baker, J. (1994). Reasoning strategies in depression: Effects of depressed mood on a syllogism task. *Personality and Individual Differences, 17*(5), 707–711.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*(4), 391–416.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18*(3), 293–328.

Chomsky, N. (1965). *Aspects of the theory of syntax.* The MIT Press.

Christie, C. (2007). Relevance theory and politeness. *Journal of Politeness Research, 3*(2), 269–294.

Cohen, J. (1973). Brief notes: Statistical power analysis and research results. *American Educational Research Journal, 10*(3), 225–229.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003.

Colling, L. J., & Szűcs, D. (2021). Statistical inference and the replication crisis. *Review of Philosophy and Psychology, 12*(1), 121–147.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science, 1*(3), Article 140216.

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of *p*-values. *Royal Society Open Science, 4*(12), Article 171085.

Comrie, B. (1986). Conditionals: A typology. In E. C. Traugott, A. Ter Meulen, J. S. Reilly, & C. A. Ferguson (Eds.), *On conditionals* (pp. 77–99). Cambridge University Press.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). Russell Sage Foundation.

Cooper, D. M., & Thompson, R. (1977). A note on the estimation of the parameters of the autoregressive-moving average process. *Biometrika, 64*(3), 625–628.

Copi, I. M. (1982). *Introduction to logic* (6th ed.). MacMillan.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*(3), 187–276.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Crowell, A., & Kuhn, D. (2014). Developing dialogic argumentation skills: A 3-year intervention study. *Journal of Cognition and Development, 15*(2), 363–381.

Cullen, A. C., & Frey, H. C. (1999). *Probabilistic techniques in exposure assessment: A handbook for dealing with variability and uncertainty in models and inputs.* Plenum Press.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286–300.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29.

Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the new statistics: estimation, open science, and beyond.* Routledge.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory and Cognition, 23*(5), 646–658.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory and Cognition, 19*(3), 274–282.

Darmstadter, H. (2013). Why do humans reason? A pragmatist supplement to an argumentative theory. *Thinking and Reasoning, 19*(4), 472–487.

Dasen, P. R. (1972). Cross-cultural Piagetian research: A summary. *Journal of Cross-Cultural Psychology, 3*(1), 23–40.

Dawkins, R., & Krebs, J. R. (1978). Animal signals: Information or manipulation? In J. R. Krebs & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (pp. 282–309). Blackwell.

De Leeuw, J., & Meijer, E. (Eds.). (2008). *Handbook of multilevel analysis.* Springer.

Demeure, V., Bonnefon, J. F., & Raufaste, E. (2009). Politeness and conditional reasoning: Interpersonal cues to the indirect suppression of deductive inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 260–266.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory and Cognition, 30*(6), 908–920.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003a). Causal conditional reasoning and strength of association: The disabling condition case. *European Journal of Cognitive Psychology, 15*(2), 161–176.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003b). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory and Cognition, 31*(4), 581–595.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking and Reasoning, 11*(4), 349–381.

Descartes, R. (1637). *Discours de la méthode.* Vrin.

Dessalles, J.-L. (2007). *Why we talk: The evolutionary origins of language.* Oxford University Press.

Dessalles, J.-L. (2011). Reasoning as a lie detection device. *Behavioral and Brain Sciences, 34*(2), 76–77.

De Vega, M., Urrutia, M., & Riffo, B. (2007). Canceling updating in the comprehension of counterfactuals embedded in narratives. *Memory and Cognition, 35*(6), 1410–1421.

Dhami, M K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*(6), 959–988.

Díaz-Pérez, F. J. (2014). Relevance theory and translation: Translating puns in Spanish film titles into English. *Journal of Pragmatics, 70*(1), 108–129.

Dibbets, P., & Meesters, C. (2020). Disconfirmation of confirmation bias: The influence of counter-attitudinal information. *Current Psychology, 41*(1), 2327–2333.

Diener, E., Northcott, R., Zyphur, M. J., & West, S. G. (2022). Beyond experiments. *Perspectives on Psychological Science.* Advance online publication. https://doi.org/10.1177/17456916211037670

Dieussaert, K., De Neys, W., & Schaeken, W. (2005). Suppression and belief revision, two sides of the same coin*? Psychologica Belgica, 45*(1), 29–46.

Dogramaci, S. (2020). What is the function of reasoning? On Mercier and Sperber's argumentative and justificatory theories. *Episteme, 17*(3), 316–330.

Douven, I., Elqayam, S., Singmann, H., & Wijnbergen-Huitink, J. V. (2020). Conditionals and inferential connections: Toward a new semantics. *Thinking and Reasoning, 26*(3), 311–351.

Drummond, C., & Fischhoff, B. (2019). Does “putting on your thinking cap” reduce myside bias in evaluation of scientific evidence? *Thinking and Reasoning, 25*(4), 477–505.

Dryzek, J. S., & Niemeyer, S. (2006). Reconciling pluralism and consensus as political ideals. *American Journal of Political Science, 50*(3), 634–649.

Durlak, J. A., & Lipsey, M. W. (1991). A practitioner's guide to meta-analysis. American *Journal of Community Psychology, 19*(3), 291–332.

Earman, J. (1992). *Bayes or bust?* The MIT Press.

Edgington, D. (1995). On conditionals. *Mind, 104*(414), 235–329.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology, 71*(1), 5–24.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* Chapman & Hall.

Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science, 21*(4), 419–460.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* Cambridge University Press.

Elqayam, S., Thompson, V. A., Wilkinson, M. R., Evans, J. S. B., & Over, D. E. (2015). Deontic introduction: A theory of inference from is to ought. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(5), 1516–1532.

Espino, O., Byrne, R. M., & Johnson-Laird, P. N. (2020). Possibilities and the parallel meanings of factual and counterfactual conditionals. *Memory and Cognition, 48*(7), 1263–1280.

Evans, J. St. B. T. (1972). Interpretation and matching bias in a reasoning task. *Quarterly Journal of Experimental Psychology, 24*(2), 193–199.

Evans, J. St. B. T. (1977). Linguistic factors in reasoning. *Quarterly Journal of Experimental Psychology, 29*(2), 297–306.

Evans, J. St. B. T. (1982). *The psychology of deductive reasoning.* Routledge & Kegan Paul.

Evans, J. St. B. T. (1983). Linguistic determinants of bias in conditional reasoning. *Quarterly Journal of Experimental Psychology, 35*(4), 635–644.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Erlbaum.

Evans, J. St. B. T. (1995). Relevance and reasoning. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning* (pp. 147–172). Erlbaum.

Evans, J. St. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology, 87*(2), 223–240.

Evans, J. St. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking and Reasoning, 4*(1), 45–110.

Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*(6), 978–996.

Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459.

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review, 13*(3), 378–395.

Evans, J. St. B. T. (2007) *Hypothetical thinking: Dual processes in reasoning and judgement.* Psychology Press.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*(1), 255–278.

Evans, J. St. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking and Reasoning, 18*(1), 5–31.

Evans, J. St. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking and Reasoning, 25*(4), 383–415.

Evans, J. St. B. T., Neilens, H., Handley, S. J., & Over, D. E. (2008). When can we say ‘if’? *Cognition, 108*(1), 100–116.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* Erlbaum.

Evans, J. St. B. T., & Over, D. E. (2004). *If.* Oxford University Press.

Evans, J. St. B. T., Over, D. E., & Handley, S. J. (2005). Suppositions, extensionality, and conditionals: A critique of the mental model theory of Johnson-Laird and Byrne (2002). *Psychological Review, 112*(4), 1040–1052.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology, 46*(4), 621–646.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33*(5), 517.

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology, 66*(1), 68–80.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge.* New Left Books.

Feyerabend, P. (1978). *Science in a free society.* New Left Books.

Fiddick, L., Brase, G. L., Cosmides, L., & Tooby, J. (2017). Rethinking relevance: Repetition priming reveals the psychological reality of adaptive specializations for reasoning. *Evolution and Human Behavior, 38*(3), 366–375.

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science, 12*(1), 46–61.

Fiedler, K. (2018). The creative cycle and the growth of psychological science. *Perspectives on Psychological Science, 13*(4), 433–438.

Fiedler, K. (2020a). Elusive alpha and beta control in a multicausal world. *Basic and Applied Social Psychology, 42*(2), 79–87.

Fiedler, K. (2020b). Cognitive representations and the predictive brain depend heavily on the environment. *Behavioral and Brain Sciences, 43*(1), Article e132.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*(6), 661–669.

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science, 16*(4), 816–826.

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—Illustrated by the report of the Open Science Collaboration. *Basic and Applied Social Psychology, 40*(3), 115–124.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665–694.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* SAGE Publications.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics, 41*(1), 155–160.

Fitelson, B., & Hitchcock, C. (2011). Probabilistic measures of causal strength. In McKay Illari, P., Russo, F., & Williamson, J. (Eds.), *Causality in the sciences* (pp. 600–627). Oxford University Press.

Fodor, J. A. (1975). *The language of thought.* Harvard University Press.

Fodor, J. A. (1983). *The modularity of mind.* The MIT Press.

Forgas, J. P., & Lantos, D. (2020). Understanding populism: Collective narcissism and the collapse of democracy in Hungary. In Forgas, J. P., Crano, W. D., & Fiedler, K. (Eds.), *Applications of social psychology: How social psychology can contribute to the solution of real-world problems* (pp. 267–291). Routledge.

Forgues, H. L., & Markovits, H. (2010). Conditional reasoning under time constraint: Information retrieval and inhibition. *Thinking and Reasoning, 16*(3), 221–232.

Fox, J., Glasspool, D., Grecu, D., Modgil, S., South, M., & Patkar, V. (2007). Argumentation-based inference and decision making—A medical perspective. *IEEE Intelligent Systems, 22*(6), 34–41.

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition, 108*(3), 819–824.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*(6084), 998.

Franken, N. (1997). Vagueness and approximation in relevance theory. *Journal of Pragmatics, 28*(2), 135–151.

Gallotti, M., & Frith, C. D. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences, 17*(4), 160–165.

Gangemi, A., Mancini, F., & Johnson-Laird, P. N. (2014). Emotion, reasoning, and psychopathology. In I. Blanchette (Ed.), *Emotion and reasoning* (pp. 44–64). Psychology Press.

Gärdenfors, P. (2003). *How homo became sapiens: On the evolution of thinking.* Oxford University Press.

Gastil, J., & Dillard, J. P. (1999). Increasing political sophistication through public deliberation. *Political Communication, 16*(1), 3–23.

Gawronski, B., & Bodenhausen, G. V. (Eds.). (2015). *Theory and explanation in social psychology.* Guilford Press.

Gawronski, B., & Strack, F. (Eds.). (2012). *Cognitive consistency: A fundamental principle in social cognition.* Guilford Press.

Gazdar, G. (1979). *Pragmatics: Implicature, presupposition, and logical form.* Academic Press.

Gazzo Castañeda, L. E., & Knauff, M. (2016a). Defeasible reasoning with legal conditionals. *Memory and Cognition, 44*(3), 499–517.

Gazzo Castañeda, L. E., & Knauff, M. (2016b). When will is not the same as should: The role of modals in reasoning with legal conditionals. *Quarterly Journal of Experimental Psychology, 69*(8), 1480–1497.

Gazzo Castañeda, L. E., & Knauff, M. (2018). Quantifying disablers in reasoning with universal and existential rules. *Thinking and Reasoning, 24*(3), 344–365.

Gazzo Castañeda, L. E., & Knauff, M. (2019). The specificity of terms affects conditional reasoning. *Thinking and Reasoning, 25*(1), 72–93.

Gazzo Castañeda, L. E., & Knauff, M. (2021a). Defeasible reasoning and belief revision in psychology. In M. Knauff & W. Spohn (Eds.), *The handbook of rationality* (pp. 347–359). The MIT Press.

Gazzo Castañeda, L. E., & Knauff, M. (2021b). Everyday reasoning with unfamiliar conditionals. *Thinking and Reasoning, 27*(3), 389–416.

Gazzo Castañeda, L. E., & Knauff, M. (2021c). Specificity effects in reasoning with counterintuitive and arbitrary conditionals. *Memory and Cognition, 50*(2), 366–377.

Gazzo Castañeda, L. E., Richter, B., & Knauff, M. (2016). Negativity bias in defeasible reasoning. *Thinking and Reasoning, 22*(2), 209–220.

Geiger, S. M., & Oberauer, K. (2007). Reasoning with conditionals: Does every counterexample count? It's frequency that counts. *Memory and Cognition, 35*(8), 2060–2074.

Gelfand, M. J., & Lorente, R. (2021). Threat, tightness, and the evolutionary appeal of populist leaders. In J. P. Forgas, W. D. Crano, & K. Fiedler (Eds.), *The psychology of populism: The tribal challenge to liberal democracy* (pp. 276–294). Routledge.

Genishi, C., & DiPaolo, M. (1982). Learning through argument in a preschool. In L. C. Wilkinson (Ed.), *Communicating in the classroom* (pp. 49–68). Academic Press.

George, C. (1997). Reasoning from uncertain premises. *Thinking and Reasoning, 3*(3), 161–189.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273–278.

Gibbard, A. (1990). *Wise choices, apt feelings: A theory of normative judgment.* Oxford University Press.

Gibbs Jr., R. W., & Tendahl, M. (2006). Cognitive effort and effects in metaphor comprehension: Relevance theory and psycholinguistics. *Mind and Language, 21*(3), 379–403.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43*(2), 127–171.

Giner-Sorolla, R. (2019). From crisis of evidence to a “crisis” of relevance? Incentive-based answers for social psychology’s perennial relevance worries. *European Review of Social Psychology, 30*(1), 1–38.

Girotto, V., Kemmelmeier, M., Sperber, D., & Van der Henst, J.-B. (2001). Inept reasoners or pragmatic virtuosos? Relevance and the deontic selection task. *Cognition, 81*(2), B69–B76.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* SAGE Publications.

Glassner, A., Weinstock, M., & Neuman, Y. (2005). Pupils' evaluation and generation of evidence and explanation in argumentation. *British Journal of Educational Psychology, 75*(1), 105–118.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*(10), 535–549.

Gonsrth, F., & Perelman, C. (1949). Philosophies premières et philosophie régressive. *Dialectica, 3*(3), 175–191.

Gorayska, B., & Lindsay, R. (1993). The roots of relevance. *Journal of Pragmatics, 19*(4), 301–323.

Grasman, R. (2017, November 15). *Meta-analysis in JASP.* JASP. https://jasp-stats.org/2017/11/15/meta-analysis-jasp

Gray, K. (2017). How to map theory: Reliable methods are fruitless without rigorous theory. *Perspectives on Psychological Science, 12*(5), 731–741.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95–112.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*(1), 1–20.

Grice, H. P. (1957). Meaning. *Philosophical Review, 66*(1), 377–388.

Grice, H. P. (1967). *Logic and conversation.* William James Lectures, Harvard University.

Grice, H. P. (1969). Utterer's meaning and intentions. *Philosophical Review, 78*(2), 147–177.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Academic Press.

Grice, H. P. (1982). Meaning revisited. In N. Smith (Ed.), *Mutual knowledge* (pp. 223–243). Academic Press.

Grice, H. P. (1989). *Studies in the way of words.* Harvard University Press.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Erlbaum.

Gutmann, A., & Thompson, D. F. (1996). *Democracy and disagreement.* Harvard University Press.

Gutt, E.-A. (2000). *Translation and relevance: Cognition and context.* St. Jerome Publishing.

Habermas, J. (1981). *Theorie des kommunikativen Handelns.* Suhrkamp.

Habermas, J. (1992). *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaates.* Suhrkamp.

Hahn, U. (2014). The Bayesian boom: Good thing or bad? *Frontiers in Psychology, 5*(1), Article 765.

Hahn, U., & Collins, P. (2021). Argumentation theory. In M. Knauff & W. Spohn (Eds.), *The handbook of rationality* (pp. 361–366). The MIT Press.

Hahn, U., & Oaksford, M. (2012). Rational argument. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 277–298). Oxford University Press.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Hall, A. (2013). Relevance theory, semantic content, and pragmatic enrichment. In A. Capone, F. Lo Piparo, & M. Carapezza (Eds.), *Perspectives on linguistic pragmatics* (pp. 99–130). Springer.

Halsey, L. G. (2019). The reign of the *p*-value is over: What alternative analyses could we employ to fill the power vacuum? *Biology Letters, 15*(5), Article 20190174.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle *P* value generates irreproducible results. *Nature Methods, 12*(3), 179–185.

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). Guilford Press.

Happé, F. G. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition, 48*(2), 101–119.

Harari, Y. N. (2014). *Sapiens: A brief history of humankind.* HarperCollins.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Wiley.

Harkness, J. A., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T.

Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 453–473). Wiley.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72*(358), 320–338.

Hasson, U., & Johnson-Laird, P. N. (2003, July 31–August 2). *Why believability cannot explain belief revision* [Paper presentation]. 25th Annual Conference of the Cognitive Science Society, Boston, MA, USA.

Hasson, U., & Walsh, C. (2003). The resolution of MP and MT inconsistencies [Unpublished manuscript]. Department of Psychology, Princeton University.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics, 17*(4), 279–296.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Academic Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83.

Hernandez, I., & Preston, J. L. (2013). Disfluency disrupts the confirmation bias. *Journal of Experimental Social Psychology, 49*(1), 178–182.

Hilton, D. J., Jaspars, J. M., & Clarke, D. D. (1990). Pragmatic conditional reasoning: Context and content effects on the interpretation of causal assertions. *Journal of Pragmatics, 14*(5), 791–812.

Hilton, D. J., Kemmelmeier, M., & Bonnefon, J. F. (2005). Putting *if*s to work: Goal-based relevance in conditional directives. *Journal of Experimental Psychology: General, 134*(3), 388–405.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39*(1), 101–117.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Holmes Finch, W., Bolin, J. E., & Kelley, K. (2014). *Multilevel modeling using R.* CRC Press.

Horn, L. R. (I984). Towards a new taxonomy for pragmatic inference. In D. Schiffrin (Ed.), *Meaning, form, and use in context* (pp. 11–42). Georgetown University Press.

Horn, L. R. (1989). *A natural history of negation.* Chicago University Press.

Horn, L. R. (2000). From if to iff: Conditional perfection as pragmatic strengthening. *Journal of Pragmatics, 32*(3), 289–326.

Horn, L. R. (2006). The border wars: A neo-Gricean perspective. In K. Turner & K. von Heusinger, *Where semantics meets pragmatics* (pp. 21–48). Elsevier.

Hornikx, J., & Hahn, U. (2012). Reasoning and argumentation: Towards an integrated psychology of argumentation. *Thinking and Reasoning, 18*(3), 225–243.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). Carus.

Huang, H., & Yang, X. (2014). Metaphor interpretation and motivation in relevance theory. *Journal of Pragmatics, 60*(1), 266–273.

Huber, W. (2016). A clash of cultures in discussions of the *P* value. *Nature Methods, 13*(8), 607.

Hughes, B. M. (2018). *Psychology in crisis.* Palgrave.

Ifantidou, E. (2014). *Pragmatic competence and relevance.* John Benjamins.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures.* Basic Books.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), Article e124.

Janis, I. L. (1972). *Victims of groupthink.* Houghton Mifflin.

Jary, M. (1998). Relevance theory and the communication of politeness. *Journal of Pragmatics, 30*(1), 1–19.

JASP Team (2022). *JASP* (Version 0.16.1) [Computer software]. https://jasp-stats.org/

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532.

Johnson, J. A. (2011). The argumentative theory of reasoning applies to scientists and philosophers, too. *Behavioral and Brain Sciences, 34*(2), 81–82.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge University Press.

Johnson-Laird, P. N. (2006) *How we reason.* Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction.* Erlbaum.

Johnson-Laird, P. N., & Byrne, R. M. J. (1994). Models, necessity, and the search for counterexamples. *Behavioral and Brain Sciences, 17*(4), 775–777.

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109*(4), 646–678.

Johnson-Laird, P. N., Byrne, R. M. J., & Girotto, V. (2009). The mental model theory of conditionals: A reply to Guy Politzer. *Topoi, 28*(1), 75–80.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review, 111*(3), 640–661.

Johnson-Laird, P. N., & Khemlani, S. S. (2013). Toward a unified theory of reasoning. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 59, pp. 1–42). Elsevier.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences, 19*(4), 201–214.

Johnson-Laird, P. N., Mancini, F., & Gangemi, A. (2006). A hyper-emotion theory of psychological illnesses. *Psychological Review, 113*(4), 822–841.

Johnson-Laird, P. N., & Oatley, K. (2000). Cognitive and social construction in emotion. In M. Lewis & J. Haviland (Eds.), *Handbook of emotions* (2nd ed., pp. 458–475). Guilford Press.

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition, 193*(1), Article 103950.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54–69.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*(1), 601–625.

Kahan, D. M. (2011). Neutral principles, motivated cognition, and some problems for constitutional law. *Harvard Law Review, 125*(1), 1–77.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697–720.

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience, 23*(1), 130–137.

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist, 86*(1), 65–79.

Kern, L. H., Mirels, H. L., & Hinshaw, V. G. (1983). Scientists' understanding of propositional logic: An experimental investigation. *Social Studies of Science, 13*(1), 131–146.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217.

Khemlani, S. S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science, 42*(6), 1887–1924.

Khemlani, S. S., Orenes, I., & Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology, 24*(5), 541–559.

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development, 71*(5), 1347–1366.

Klauer, K. C., Beller, S., & Hütter, M. (2010). Conditional reasoning in context: A dual-source model of probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(2), 298–323.

Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 680–703.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*(2), 211–228.

Knauff, M., & Gazzo Castañeda, L. E. (2021). When nomenclature matters: Is the “new paradigm” really a new paradigm for the psychology of reasoning? *Thinking and Reasoning,* 1–30. https://doi.org/10.1080/13546783.2021.1990126

Knauff, M., & Knoblich, G. (2017). Logisches Denken. In J. Müsseler & M. Rieger (Eds.), *Allgemeine Psychologie* (3rd ed., pp. 533–585). Springer.

Knauff, M., & Ragni, M. (2011). Cross-cultural preferences in spatial reasoning. *Journal of Cognition and Culture, 11*(1), 1–21.

Knauff, M., & Spohn, W. (Eds.). (2021). *The handbook of rationality.* The MIT Press.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation? In J. R. Krebs & N. B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (2nd ed., pp. 390–402). Basil Blackwell.

Kreft, I. G., & de Leeuw, J. (1998). *Introducing multilevel modeling.* SAGE Publications.

Krzyżanowska, K., Collins, P. J., & Hahn, U. (2017). Between a conditional’s antecedent and its consequent: Discourse coherence vs. probabilistic relevance. *Cognition, 164*(1), 199–205.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*(4), 674–689.

Kuhn, D. (1991). *The skills of argument.* Cambridge University Press.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education, 77*(3), 319–337.

Kuhn, D. (2001). How do people know? *Psychological Science, 12*(1), 1–8.

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development, 15*(3), 309–328.

Kuhn, D., & Crowell, A. (2011). Dialogic argumentation as a vehicle for developing young adolescents’ thinking. *Psychological Science, 22*(4), 545–552.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development, 74*(5), 1245–1260.

Kuhn, D., Wang, Y., & Li, H. (2010). Why argue? Developing understanding of the purposes and values of argumentative discourse. *Discourse Processes, 48*(1), 26–49.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–195). Cambridge University Press.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*(1), Article 863.

Landemore, H., & Mercier, H. (2012). Talking it out with others vs. deliberation within and the law of group polarization: Some implications of the argumentative theory of reasoning for deliberative democracy. *Análise Social, 47*(205), 910–934.

Laughlin, P. R. (2011). *Group problem solving.* Princeton University Press.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology, 22*(3), 177–189.

Lazzeroni, L. C., Lu, Y., & Belitskaya-Lévy, I. (2016). Solutions for quantifying *P*-value uncertainty and replication power. *Nature Methods, 13*(2), 107–108.

Ledgerwood, A., Soderberg, C. K., & Sparks, J. (2017). Designing a study to maximize informational value. In M. C. Makel & J. A. Plucker (Eds.), *Toward a more perfect psychology: Improving trust, accuracy, and transparency in research* (pp. 33–58). American Psychological Association.

Levinson, S. C. (1983). *Pragmatics.* Cambridge University Press.

Levinson, S. C. (1989). A review of relevance. *Journal of Linguistics, 25*(2), 455–472.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* The MIT Press.

Levinson, S. C., Kita, S., Haun, D. B., & Rasch, B. H. (2002). Returning the tables: Language affects spatial reasoning. *Cognition, 84*(2), 155–188.

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications, 11*(1), 1–12.

Lewin, K. (1943). Psychology and the process of group living. *The Journal of Social Psychology, 17*(1), 113–131.

Li, P., & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition, 83*(3), 265–294.

Lindenberg, S. (2001). Social rationality as a unified model of man (including bounded rationality). *Journal of Management and Governance, 5*(3), 239–251.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181–1209.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* SAGE Publications.

Liu, L., Li, C., & Zhu, D. (2012). A new approach to testing nomological validity and its application to a second-order measurement model of trust. *Journal of the Association for Information Systems, 13*(12), 950–975.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6*(1), Article 1171.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584–585.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098–2109.

Lucas, E., & Ball, L. (2005). Think-aloud protocols and the selection task: Evidence for relevance effects and rationalisation processes. *Thinking and Reasoning, 11*(1), 35–66.

Lucy, J. A. (1992a). *Grammatical categories and cognition: A case study of the linguistic relativity hypothesis.* Cambridge University Press.

Lucy, J. A. (1992b). *Language diversity and thought: A reformulation of the linguistic relativity hypothesis.* Cambridge University Press.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92.

Mandel, D. R., Hilton, D. J., & Catellani, P. E. (2005). *The psychology of counterfactual thinking.* Routledge.

Mandelbaum, D. G. (Ed.). (1951). *Selected writings of Edward Sapir in language, culture, and personality.* University of California Press.

Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science, 9*(3), 343–351.

Manktelow, K. I. (2012). *Thinking and reasoning: An introduction to the psychology of reason, judgment and decision making.* Psychology Press.

Manktelow, K. I., & Fairley, N. (2000). Superordinate principles in reasoning with causal and deontic conditionals. *Thinking and Reasoning, 6*(1), 41–65.

Manktelow, K. I., Fairley, N., Kilpatrick, S. G., & Over, D. E. (2000). Pragmatics and strategies for practical reasoning. In W. Schaeken, G. De Vooght, A.

Vandierendonk, & G. d'Ydewalle (Eds.), *Deductive reasoning and strategies* (pp. 111–130). Erlbaum.

Marcus, S. L., & Rips, L. J. (1979). Conditional reasoning. *Journal of Verbal Learning and Verbal Behavior, 18*(2), 199–223.

Markman, K. D., McMullen, M. N., & Elizaga, R. A. (2008). Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. *Journal of Experimental Social Psychology, 44*(2), 421–428.

Markovits, H. (1986). Familiarity effects in conditional reasoning. *Journal of Educational Psychology, 78*(6), 492–494.

Markovits, H. (1988). Conditional reasoning, representation, and empirical evidence on a concrete task. *The Quarterly Journal of Experimental Psychology, 40*(3), 483–495.

Markovits, H., Fleury, M. L., Quinn, S., & Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child Development, 69*(3), 742–755.

Markovits, H., Forgues, H. L., & Brunet, M. L. (2010). Conditional reasoning, frequency of counterexamples, and the effect of response modality. *Memory and Cognition, 38*(4), 485–492.

Markovits, H., & Potvin, F. (2001). Suppression of valid inferences and knowledge structures: The curious effect of producing alternative antecedents on reasoning with causal conditionals. *Memory and Cognition, 29*(5), 736–744.

Markovits, H., & Quinn, S. (2002). Efficiency of retrieval correlates with "logical" reasoning from causal conditional premises. *Memory and Cognition, 30*(5), 696–706.

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.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*(1), 305–315.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*(1), 537–563.

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.

Mayo, D. G. (2021). Significance tests: Vitiated or vindicated by the replication crisis in psychology? *Review of Philosophy and Psychology, 12*(1), 101–120.

Mazzarella, D. (2015). Politeness, relevance and scalar inferences. *Journal of Pragmatics, 79*(1), 93–106.

McNemar, Q. (1960). At random: Sense and nonsense. *American Psychologist, 15*(5), 295–300.

Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction, 14*(2), 139–178.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review, 10*(3), 517–532.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834.

Melton, R. J. (1995). The role of positive affect in syllogism performance. *Personality and Social Psychology Bulletin, 21*(8), 788–794.

Mercier, H. (2011a). On the universality of argumentative reasoning. *Journal of Cognition and Culture, 11*(1), 85–113.

Mercier, H. (2011b). What good is moral reasoning? *Mind and Society, 10*(2), 131–148.

Mercier, H. (2011c). When experts argue: Explaining the best and the worst of reasoning. *Argumentation, 25*(3), 313–327.

Mercier, H. (2011d). Reasoning serves argumentation in children. *Cognitive Development, 26*(3), 177–191.

Mercier, H. (2013a). The function of reasoning: Argumentative and pragmatic alternatives. *Thinking and Reasoning, 19*(4), 488–494.

Mercier, H. (2013b). Using evolutionary thinking to cut across disciplines: The example of the argumentative theory of reasoning. In P. H. Crowley & T. R. Zentall (Eds.), *Comparative decision making* (pp. 279–304). Oxford University Press.

Mercier, H. (2016a). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences, 20*(9), 689–700.

Mercier, H. (2016b). Making science education more natural—Some ideas from the argumentative theory of reasoning. *Zeitschrift für Pädagogische Psychologie, 30*(2–3), 151–153.

Mercier, H., Boudry, M., Paglieri, F., & Trouche, E. (2017). Natural-born arguers: Teaching how to make the best of our reasoning abilities. *Educational Psychologist, 52*(1), 1–16.

Mercier, H., Deguchi, M., Van der Henst, J. B., & Yama, H. (2016). The benefits of argumentation are cross-culturally robust: The case of Japan. *Thinking and Reasoning, 22*(1), 1–15.

Mercier, H., & Heintz, C. (2014). Scientists' argumentative reasoning. *Topoi, 33*(2), 513–524.

Mercier, H., & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology, 33*(2), 243–258.

Mercier, H., & Sperber, D. (2009). Intuitive and reflective inferences. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual Processes and beyond* (pp. 149–170). Oxford University Press.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74.

Mercier, H., & Sperber, D. (2017). *The enigma of reason.* Harvard University Press.

Miller, G. A., & Beebe-Center, J. G. (1956). Some psychological methods for evaluating the quality of translations. *Mechanical Translation, 3*(1), 73–80.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*(4), 201–218.

Morís Fernández, L., & Vadillo, M. A. (2020). Flexibility in reaction time analysis: Many roads to a false positive? *Royal Society Open Science, 7*(2), Article 190831.

Morris, C. (1938). *Foundations of the theory of signs.* University of Chicago Press.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking and Reasoning, 4*(3), 231–248.

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science, 28*(2), 135–167.

Myers, D. G., & Lamm, H. (1975). The polarizing effect of group discussion: The discovery that discussion tends to enhance the average prediscussion tendency has stimulated new insights about the nature of group influence. *American Scientist, 63*(3), 297–303.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2*(2), 842–860.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.

Nicolle, S. (1998). A relevance theory perspective on grammaticalization. *Cognitive Linguistics, 9*(1), 1–35.

Niemeyer, S., & Dryzek, J. S. (2007). The ends of deliberation: Meta-consensus and inter-subjective rationality as ideal outcomes. *Swiss Political Science Review, 13*(4), 497–526.

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently ... and why.* Free Press.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291–310.

Noh, E. J. (1996). A relevance-theoretic account of metarepresentative uses in conditionals. *UCL Working Papers in Linguistics, 8*(1), 125–163.

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science, 26*(5), 653–684.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education, 87*(2), 224–240.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology, 73*(1), 719–748.

Nussbaum, E. M. (2008). Using argumentation vee diagrams (AVDs) for promoting argument–counterargument integration in reflective writing. *Journal of Educational Psychology, 100*(3), 549–565.

Nussbaum, E. M., & Asterhan, C. S. C. (2016). The psychology of far transfer from classroom argumentation. In F. Paglieri, L. Bonelli, & S. Felletti (Eds.), *The psychology of argument: Cognitive approaches to argumentation and persuasion* (pp. 407–422). College Publications.

Nuzzo, R. (2014). Statistical errors: *p* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature, 506*(7487), 150–153.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*(4), 608–631.

Oaksford, M., & Chater, N. (1995). Information gain explains relevance which explains the selection task. *Cognition, 57*(1), 97–108.

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review, 103*(2), 381–391.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin and Review, 10*(2), 289–318.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford University Press.

Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences, 32*(1), 69–84.

Oaksford, M., & Chater, N. (Eds.). (2010). *Cognition and conditionals: Probability and logic in human thinking.* Oxford University Press.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology, 71*(1), 305–330.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(4), 883–899.

Oaksford, M., & Hahn, U. (2007). Induction, deduction, and argument strength in human reasoning and argumentation. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches* (pp. 269–301). Cambridge University Press.

Oaksford, M., Morris, F., Grainger, B., & Williams, J. M. G. (1996). Mood, reasoning, and central executive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(2), 476–492.

Oberauer, K., & Greve, W. (2021). Intentional remembering and intentional forgetting in working and long-term memory. *Journal of Experimental Psychology: General.* Advance online publication. https://doi.org/10.1037/xge0001106

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin and Review, 26*(5), 1596–1618.

Oberauer, K., Weidenfeld, A., & Fischer, K. (2007). What makes us believe a conditional? The roles of covariation and causality. *Thinking and Reasoning, 13*(4), 340–369.

O'Brien, D. P. (2009). Human reasoning includes a mental logic. *Behavioral and Brain Sciences, 32*(1), 96–97.

O'Neill, T. P. (2012). Law and the Argumentative Theory. *Oregon Law Review, 90*(3), 837–854.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943–951.

Osgood, C. E. (1962). Studies on the generality of affective meaning systems. *American Psychologist, 17*(1), 10–28.

Osgood, C. E. (1969). On the whys and wherefores of E, P, and A. *Journal of Personality and Social Psychology, 12*(3), 194–199.

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning.* University of Illinois Press.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning.* University of Illinois Press.

Oswald, S. (2007). Towards an interface between pragma-dialectics and relevance theory. *Pragmatics and Cognition, 15*(1), 179–201.

Over, D. E., Hadjichristidis, C., Evans, J. S. B., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology, 54*(1), 62–97.

Oyserman, D. (2017). Culture three ways: Culture and subcultures within countries. *Annual Review of Psychology, 68*(1), 435–463.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58*(3), 545–554.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge University Press.

Peloquin, B. N., Goodman, N. D., & Frank, M. C. (2020). The interactions of rational, pragmatic agents lead to efficient language structure and use. *Topics in Cognitive Science, 12*(1), 433–445.

Peng, K., & Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American Psychologist, 54*(9), 741–754.

Pennington, N., & Hastie, R. (1981). Juror decision-making models: The generalization gap. *Psychological Bulletin, 89*(2), 246–287.

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology, 51*(2), 242–258.

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 521–533.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology, 62*(2), 189–206.

Pennington, N., & Hastie, R. (1993). Reasoning in explanation-based decision-making. *Cognition, 49*(1–2), 123–163.

Perelman, C., & Olbrechts-Tyteca, L. (1969). *The new rhetoric: A treatise on argumentation.* University of Notre Dame Press.

Pérez Zafrilla, P. J. (2016). Is deliberative democracy an adaptive political theory? A critical analysis of Hugo Mercier's argumentative theory of reasoning. *Análise Social, 220*(51), 544–564.

Pfeifer, N. (2013). The new psychology of reasoning: A mental probability logical perspective. *Thinking and Reasoning, 19*(3), 329–345.

Pfeifer, N., & Kleiter, G. D. (2010). The conditional in mental probability logic. In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinking* (pp. 153–173). Oxford University Press.

Piaget, J. (1928). *Judgment and reasoning in the child.* Routledge.

Piaget, J., & Inhelder, B. (1974). *The child's construction of quantities: Conservation and atomism.* Routledge.

Pilkington, A. (2000). *Poetic effects: A relevance theory perspective.* John Benjamins.

Pinker, S. (1994). *The language instinct: How the mind creates language.* Harper Collins.

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, 146*(3642), 347–353.

Politzer, G., & Bourmaud, G. (2002). Deductive reasoning from uncertain conditionals. *British Journal of Psychology, 93*(3), 345–381.

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science, 11*(4), 481–518.

Pollock, J. L. (2001). Defeasible reasoning with variable degrees of justification. *Artificial Intelligence, 133*(1–2), 233–282.

Popper, K. (1935). *Logik der Forschung.* Springer.

Popper, K. (1945). *The open society and its enemies.* Routledge.

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge.* Routledge.

Popper, K. (1972). *Objective knowledge: An evolutionary approach.* Oxford University Press.

Popper, K. (1976). *Unended quest: An intellectual autobiography.* Routledge.

Popper, K. (1994). *The myth of the framework: In defense of science and rationality.* Routledge.

Prado, J., Léone, J., Epinat-Duclos, J., Trouche, E., & Mercier, H. (2020). The neural bases of argumentative reasoning. *Brain and Language, 208*(1), Article 104827.

Prakken, H. (2008). AI & law on legal argument: Research trends and application prospects. *SCRIPTed, 5*(3), 449–454.

Preckel, F., & Brunner, M. (2017). Nomological nets. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–4). Springer.

Putnam, H. (1981). *Reason, truth and history.* Cambridge University Press.

Putnam, H. (2008). *Jewish philosophy as a guide to life.* Indiana University Press.

Pyszczynski, T. A., Greenberg, J., Mack, D., & Wrightsman, L. S. (1981). Opening statements in a jury trial: The effect of promising more than the evidence can show. *Journal of Applied Social Psychology, 11*(5), 434–444.

Pyszczynski, T. A., & Wrightsman, L. S. (1981). The effects of opening statements on mock jurors' verdicts in a simulated criminal trial. *Journal of Applied Social Psychology, 11*(4), 301–313.

Quelhas, A. C., & Byrne, R. M. J. (2003). Reasoning with deontic and counterfactual conditionals. *Thinking and Reasoning, 9*(1), 43–65.

Quinn, S., & Markovits, H. (1998). Conditional reasoning, causality, and the structure of semantic memory: Strength of association as a predictive factor for content effects. *Cognition, 68*(3), B93–B101.

R Core Team (2022). *R: A language and environment for statistical computing*. (Version 4.2.0) [Computer software]. https://www.R-project.org/.

Radenhausen, R. A., & Anker, J. M. (1988). Effects of depressed mood induction on reasoning performance. *Perceptual and Motor Skills, 66*(3), 855–860.

Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses: A theory of selection tasks. *Psychological Bulletin, 144*(8), 779–796.

Rahwan, I., & Moraitis, P. (Eds.). (2009). *Argumentation in multi-agent systems.* Springer.

Rahwan, I., Zablith, F., & Reed, C. (2007). Laying the foundations for a world wide argument web. *Artificial Intelligence, 171*(10–15), 897–921.

Rajsic, J., Wilson, D. E., & Pratt, J. (2015). Confirmation bias in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 41*(5), 1353–1364.

Ramsey, F. P. (1929/1990). General propositions and causality. In D. H. Mellor (Ed.), *Philosophical papers* (pp. 145–163). Cambridge University Press.

Ransom, K. J., Perfors, A., & Navarro, D. J. (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science, 40*(7), 1775–1796.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*(3), 510–532.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* SAGE Publications.

Reed, C., & Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools, 13*(4), 961–979.

Rescher, N. (1976). *Plausible reasoning.* Van Gorcum.

Resnick, L. B., Asterhan, C. S., & Clarke, S. N. (2013). *Socializing intelligence through academic talk and dialogue.* American Educational Research Association.

Revlin, R., Cate, C. L., & Rouss, T. S. (2001). Reasoning counterfactually: Combining and rending. *Memory and Cognition, 29*(8), 1196–1208.

Richter, B. (2021). Viewing attitude research through a Brunswikian lens. *The Brunswik Society Newsletter, 36*(1), 52–54.

Richter, B., & Hütter, M. (2021). Learning of affective meaning: Revealing effects of stimulus pairing and stimulus exposure. *Cognition and Emotion, 35*(8), 1588–1606.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking.* The MIT Press.

Roberge, J. J. (1978). Linguistic and psychometric factors in propositional reasoning. *The Quarterly Journal of Experimental Psychology, 30*(4), 705–716.

Rollwage, M., & Fleming, S. M. (2021). Confirmation bias is adaptive when coupled with efficient metacognition. *Philosophical Transactions of the Royal Society B, 376*(1822), Article 20200131.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 1*(3), 377–386.

Rumain, B., Connell, J., & Braine, M. D. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: If is not the biconditional. *Developmental Psychology, 19*(4), 471–481.

Ruytenbeek, N. (2019). Indirect requests, relevance, and politeness. *Journal of Pragmatics, 142*(1), 78–89.

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching, 41*(5), 513–536.

Santamaría, C., Espino, O., & Byrne, R. M. J. (2005). Counterfactual and semifactual conditionals prime alternative possibilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1149–1154.

Sapir, E. (1921). *Language: An introduction to the study of speech.* Harcourt, Brace & Co.

Schaeken, W., Van der Henst, J., & Schroyens, W. (2007). The mental models theory of relational reasoning: Premises' relevance, conclusions' phrasing, and cognitive economy. In W. Schaeken, A. Vandierendonck, W. Schroyens, & G. d'Ydewalle (Eds.), *The mental models theory of reasoning: Refinements and extensions* (pp. 129–150). Erlbaum.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*(5), 529–540.

Schneiderhan, E., & Khan, S. (2008). Reasons and inclusion: The foundation of deliberation. *Sociological Theory, 26*(1), 1–24.

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature, 515*(7525), 9.

Schourup, L. (2011). The discourse marker now: A relevance-theoretic approach. *Journal of Pragmatics, 43*(8), 2110–2129.

Schramm, P., & Rouder, J. (2019). *Are reaction time transformations really beneficial?* PsyArXiv. https://doi.org/10.31234/osf.io/9ksa6

Schroyens, W. J., Schaeken, W., & d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking and Reasoning, 7*(2), 121–172.

Sebanz, N., & Knoblich, G. (2021). Progress in joint-action research. *Current Directions in Psychological Science, 30*(2), 138–143.

Sedlmeier, P., & Gigerenzer, G. (1992). Do studies of statistical power have an effect on the power of studies? In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 389–406). American Psychological Association.

Shadish, W. R., Cook, T. D., & Campbell D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin.

Shehab, H. M., & Nussbaum, E. M. (2015). Cognitive load of critical thinking strategies. *Learning and Instruction, 35*(1), 51–61.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*(1), 487–510.

Sikorski, M., van Dongen, N., & Sprenger, J. (2019). *Causal conditionals, tendency causal claims and statistical relevance.* PsyArXiv. https://doi.org/10.31234/osf.io/t3fud

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., ... & Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Simon, H. A. (1957). *Models of man: Social and rational.* Wiley.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76–80.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour, 4*(11), 1208–1214.

Sinaiko, H. W., & Brislin, R. W. (1973). Evaluating language translations: Experiments on three assessment methods. *Journal of Applied Psychology, 57*(3), 328–334.

Singmann, H., & Kellen, D. (2020). An introduction to linear mixed modeling in experimental psychology. In D. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (pp. 4–31). Routledge.

Singmann, H., Klauer, K. C., & Beller, S. (2016). Probabilistic conditional reasoning: Disentangling form and content with the dual-source model. *Cognitive Psychology, 88*(1), 61–87.

Singmann, H., Klauer, K. C., & Over, D. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology, 5*(1), Article 316.

Skinner, B. F. (1948). *Walden two.* Prentice Hall.

Skovgaard-Olsen, N. (2016a). Ranking theory and conditional reasoning. *Cognitive Science, 40*(4), 848–880.

Skovgaard-Olsen, N. (2016b). Motivating the relevance approach to conditionals. *Mind and Language, 31*(5), 555–579.

Skovgaard-Olsen, N., & Collins, P. (2021). Indicatives, subjunctives, and the falsity of the antecedent. *Cognitive Science, 45*(11), Article e13058.

Skovgaard-Olsen, N., Collins, P., Krzyżanowska, K., Hahn, U., & Klauer, K. C. (2019). Cancellation, negation, and rejection. *Cognitive Psychology, 108*(1), 42–71.

Skovgaard-Olsen, N., Kellen, D., Hahn, U., & Klauer, K. C. (2019). Norm conflicts and conditionals. *Psychological Review, 126*(5), 611–633.

Skovgaard-Olsen, N., Kellen, D., Krahl, H., & Klauer, K. C. (2017). Relevance differently affects the truth, acceptability, and probability evaluations of “and”, “but”, “therefore”, and “if–then”. *Thinking and Reasoning, 23*(4), 449–482.

Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition, 150*(1), 26–36.

Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2017). Relevance and reason relations. *Cognitive Science, 41*(1), 1202–1215.

Slavin, R. E. (1995). *Cooperative learning: Theory, research, and practice* (2nd ed.). Allyn and Bacon.

Smith, M. L. (1980). Publication bias and meta-analysis. *Evaluation in Education, 4*(1), 22–24.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science, 323*(5910), 122–124.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE Publications.

Spelke, E., & Tsivkin, S. (2001). Initial knowledge and conceptual change: Space and number. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 70–100). Cambridge University Press.

Spellman, B. A. (2012). Introduction to the special section: Data, data, everywhere... especially in my file drawer. *Perspectives on Psychological Science, 7*(1), 58–59.

Sperber, D. (2001). An evolutionary perspective on testimony and argumentation. *Philosophical Topics, 29*(1–2), 401–413.

Sperber, D. (2013). Speakers are honest because hearers are vigilant: Reply to Kourken Michaelian. *Episteme, 10*(1), 61–71.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 57*(1), 31–95.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language, 25*(4), 359–393.

Sperber, D., & Girotto, V. (2003). Does the selection task detect cheater-detection? In K. Sterelny & J. Fitness (Eds.), *From mating to mentality: Evaluating evolutionary psychology* (pp. 197–226) Psychology Press.

Sperber, D., & Mercier, H. (2012). Reasoning as a social competence. In H. Landemore & J. Elster (Eds.), *Collective wisdom: Principles and mechanisms* (pp. 368–392). Cambridge University Press.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition.* Basil Blackwell.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell Publishing.

Sperber, D., & Wilson, D. (1996). Fodor's frame problem and relevance theory. *Behavioral and Brain Sciences, 19*(3), 530–532.

Sperber, D., & Wilson, D. (1997). Remarks on relevance theory and the social sciences. *Multilingua, 16*(2–3), 145–151.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind and Language, 17*(1–2), 3–23.

Spohn, W. (2012). *The laws of belief: Ranking theory and its philosophical applications.* Oxford University Press.

Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science, 37*(6), 1074–1106.

Spohn, W. (2020). Defeasible normative reasoning. *Synthese, 197*(4), 1391–1428.

Sprenger, J. (2018). Foundations of a probabilistic theory of causal strength. *Philosophical Review, 127*(3), 371–398.

Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science.* Oxford University Press.

Stanovich, K. E. (2004). *The robot's rebellion.* Chicago University Press.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought.* Yale University Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind.* Oxford University Press.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking and Reasoning, 13*(3), 225–247.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*(4), 672–695.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science, 22*(4), 259–264.

Stasson, M. F., Kameda, T., Parks, C. D., Zimmerman, S. K., & Davis, J. H. (1991). Effects of assigned group consensus requirement on group problem solving and group members' learning. *Social Psychology Quarterly, 54*(1), 25–35.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712.

Steenbergen, M. R., Bächtiger, A., Spörndli, M., & Steiner, J. (2003). Measuring political deliberation: A discourse quality index. *Comparative European Politics, 1*(1), 21–48.

Stein, N. L., & Albro, E. R. (2001). The origins and nature of arguments: Studies in conflict understanding, emotion, and negotiation. *Discourse Processes, 32*(2–3), 113–133.

Stein, N. L., & Bernas, R. (1999). The early emergence of argumentative knowledge and skill. In J. Andriessen & P. Coirier (Eds.), *Foundations of argumentative text processing* (pp. 97–116). Amsterdam University Press.

Stein, N. L., & Miller, C. A. (1993). The development of meaning and reasoning skill in argumentative contexts: Evaluating, explaining, and generating evidence. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 4, pp. 285–335). Erlbaum.

Stenning, K., & van Lambalgen, M. (2004). A little logic goes a long way: Basing experiment on semantic theory in the cognitive science of conditional reasoning. *Cognitive Science, 28*(4), 481–529.

Stenning, K., & van Lambalgen, M. (2005). Semantic interpretation as computation in nonmonotonic logic: The real meaning of the suppression task. *Cognitive Science, 29*(6), 919–960.

Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science.* The MIT Press.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54*(285), 30–34.

Sterrett, J. (2012). Arguing against the argumentative theory of reasoning. *Cogency: Journal of Reasoning and Argumentation, 4*(1), 185–199.

Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *The Quarterly Journal of Experimental Psychology, 48*(3), 613–643.

Stevenson, R. J., & Over, D. E. (2001). Reasoning from uncertain premises: Effects of expertise and conversational context. *Thinking and Reasoning, 7*(4), 367–390.

Suiter, J., Farrell, D. M., Harris, C., & Murphy, P. (2021). Measuring epistemic deliberation on polarized issues: The case of abortion provision in Ireland. *Political Studies Review.* Advance online publication. https://doi.org/10.1177/14789299211020909

Suppes, P. (1966). Probabilistic inference and the concept of total evidence. In J. Hintikka & P. Suppes (Eds.), *Aspects of inductive logic* (pp. 49–65). North-Holland.

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science, 16*(4), 717–724.

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*(3), Article e2000797.

Tanaka, K. (1994). *Advertising language: A pragmatic approach to advertisements in Britain and Japan.* Routledge.

Taplin, J. E. (1971). Reasoning with conditional sentences. *Journal of Verbal Learning and Verbal Behavior, 10*(3), 219–225.

Tendahl, M., & Gibbs Jr., R. W. (2008). Complementary perspectives on metaphor: Cognitive linguistics and relevance theory. *Journal of Pragmatics, 40*(11), 1823–1864.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309–318.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition, 22*(6), 742–758.

Thompson, V. A. (1995). Conditional reasoning: The necessary and sufficient conditions. *Canadian Journal of Experimental Psychology, 49*(1), 1–58.

Thompson, V. A. (2000). The task-specific nature of domain-general reasoning. *Cognition, 76*(3), 209–268.

Thompson, D. F. (2008). Deliberative democratic theory and empirical political science. *Annual Review of Political Science, 11*(1), 497–520.

Thompson, V. A., Evans, J. S. B., & Handley, S. J. (2005). Persuading and dissuading by conditional argument. *Journal of Memory and Language, 53*(2), 238–257.

Toms, M., Morris, N., & Ward, D. (1993). Working memory and conditional reasoning. *The Quarterly Journal of Experimental Psychology, 46*(4), 679–699.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge University Press.

Trafimow, D. (2015). Rational actor theories. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 245–265). Guilford Press.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General, 143*(5), 1958–1971.

Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesley.

Twisk, J. W. (2006). *Applied multilevel analysis: A practical guide for medical researchers.* Cambridge University Press.

Tyupa, S. (2011). A theoretical framework for back-translation as a quality assessment tool. *New Voices in Translation Studies, 7*(1), 35–46.

Vadeboncoeur, I., & Markovits, H. (1999). The effect of instructions and information retrieval on accepting the premises in a conditional reasoning task. *Thinking and Reasoning, 5*(2), 97–113.

Valiña, M. D., Seoane, G., Ferraces, M. J., & Martín, M. (1999). The importance of pragmatic aspects in conditional reasoning. *The Spanish Journal of Psychology, 2*(1), 20–31.

Van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17–45). Cambridge University Press.

Vedejová, D., & Čavojová, V. (2022). Confirmation bias in information search, interpretation, and memory recall: Evidence from reasoning about four controversial topics. *Thinking and Reasoning, 28*(1), 1–28.

Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics, 32*(2), 227–230.

Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking and Reasoning, 11*(3), 239–278.

Viechtbauer, W. (2010a). *Metafor: Meta-analysis package for R* (Version 3.4-0) [Computer software]. https://CRAN.R-project.org/package=metafor

Viechtbauer, W. (2010b). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125.

Voss, J. F., & Van Dyke, J. A. (2001). Narrative structure, information certainty, emotional content, and gender as factors in a pseudo jury decision-making task. *Discourse Processes, 32*(2–3), 215–243.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274–290.

Vygotsky, L. S. (1962). *Thought and language.* The MIT Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Wachter, K. W., & Straf, M. L. (Eds.). (1990). *The future of meta-analysis.* Russell Sage Foundation.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi:

Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426–432.

Walton, D. (1989). *A handbook for critical argumentation.* Cambridge University Press.

Walton, D. (2008). *Informal logic: A pragmatic approach.* Cambridge University Press.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes.* Cambridge University Press.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology, 12*(3), 129–140.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 106–137). Penguin.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology, 20*(3), 273–281.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* Harvard University Press.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133.

Wearing, C. J. (2015). Relevance theory: Pragmatics and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(2), 87–95.

Werner, O. & Campbell, D. T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), *A handbook of method in cultural anthropology* (pp. 398–420). American Museum of Natural History.

West, B. T., Welch, K. B., & Gałecki, A. T. (2015). *Linear mixed models: A practical guide using statistical software.* CRC Press.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045.

Wharton, T., Bonard, C., Dukes, D., Sander, D., & Oswald, S. (2021). Relevance and emotion. *Journal of Pragmatics, 181*(1), 259–269.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record, 58*(3), 475–482.

Whorf, B. L. (1956). *Language, thought, and reality.* The MIT Press.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics, 9*(1), 60–62.

Williams, D. R., Mulder, J., Rouder, J. N., & Rast, P. (2021). Beneath the surface: Unearthing within-person variability and mean relations with Bayesian mixed models. *Psychological Methods, 26*(1), 74–89.

Williamson, T. (2007) Philosophical knowledge and knowledge of counterfactuals. In C. Beyer & A Burri (Eds.), *Philosophical knowledge: Its possibility and scope* (pp. 89–124). Rodopi.

Wilson, D., & Sperber, D. (1994). Outline of relevance theory. *Links and Letters, 1*(1), 85–106.

Wilson, D., & Sperber, D. (2002). Truthfulness and relevance. *Mind, 111*(443), 583–632.

Wilson, D., & Sperber, D. (Eds.). (2012a). *Meaning and relevance.* Cambridge University Press.

Wilson, D., Sperber, D. (2012b) Explaining irony. In D. Wilson & D. Sperber (Eds.), *Meaning and relevance* (pp. 123–145). Cambridge University Press.

Wittgenstein, L. (1922). *Tractatus logico-philosophicus.* Routledge & Kegan Paul.

Wittgenstein, L. (1953). *Philosophical investigations.* Macmillan.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18*(4), 451–460.

World Medical Association (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA, 310*(20), 2191–2194.

Yuan, W., Lin, F. Y., & Cooper, R. P. (2019). Relevance theory, pragmatic inference and cognitive architecture. *Philosophical Psychology, 32*(1), 98–122.

Yus, F. (1998). A decade of relevance theory. *Journal of Pragmatics, 30*(3), 305–345.

Yus, F. (2003). Humor and the search for relevance. *Journal of Pragmatics, 35*(9), 1295–1331.

Yus, F. (2010). Relevance theory. In A. Barber & R. J. Stainton (Eds.), *Concise encyclopedia of philosophy of language and linguistics* (pp. 648–655). Elsevier.

Yus, F. (2016). *Humor and relevance.* John Benjamins.

Zavala, J., & Kuhn, D. (2017). Solitary discourse is a productive activity. *Psychological Science, 28*(5), 578–586.

List of Equations

List of Figures

List of Tables

Appendix

Appendix A1: Instructions of Experiment 1

Instruction 1

Lieber Teilnehmer, liebe Teilnehmerin,

im Folgenden werden Ihnen Aufgaben präsentiert, die sich aus mehreren Aussagen zusammensetzen. Die ersten Aussagen sind in schwarzer Schrift geschrieben und beinhalten eine Wenn-dann-Aussage, einen Fakt und gegebenenfalls eine Zusatzinformation. Die letzte Aussage ist in rot geschrieben und ist eine Frage, die nach einer Schlussfolgerung fragt.

Ihre Aufgabe besteht darin, diese Frage zu beantworten. Dazu stehen Ihnen Antworten von „nein, auf keinen Fall“ bis „ja, auf jeden Fall“ zur Verfügung, die Sie mit Hilfe des Ziffernblocks angeben. Die Nummer 1 steht hierbei für „nein, auf keinen Fall“ und 7 für „ja, auf jeden Fall“.

Antworten Sie bitte so, wie Sie es auch in Alltagssituationen machen würden.

Instruction 1 (counterbalanced)

Lieber Teilnehmer, liebe Teilnehmerin,

im Folgenden werden Ihnen Aufgaben präsentiert, die sich aus mehreren Aussagen zusammensetzen. Die ersten Aussagen sind in schwarzer Schrift geschrieben und beinhalten eine Wenn-dann-Aussage, einen Fakt und gegebenenfalls eine Zusatzinformation. Die letzte Aussage ist in rot geschrieben und ist eine Frage, die nach einer Schlussfolgerung fragt.

Ihre Aufgabe besteht darin, diese Frage zu beantworten. Dazu stehen Ihnen Antworten von „ja, auf jeden Fall“ bis „nein, auf keinen Fall“ zur Verfügung, die Sie mit Hilfe des

Ziffernblocks angeben. Die Nummer 1 steht hierbei für „ja, auf jeden Fall“ und 7 für „nein, auf keinen Fall“.

Antworten Sie bitte so, wie Sie es auch in Alltagssituationen machen würden.

Instruction 2

Von Aussage zu Aussage kommen Sie immer mit der Leertaste weiter und zwischen den Aufgaben wird Ihnen die Möglichkeit gegeben, eine Pause zu machen. Die Pause beenden Sie ebenfalls mit der Leertaste. Lesen Sie bitte jede Aufgabe sorgfältig durch, da die Aussagen auch Verneinungen enthalten können.

Alles klar? Wenn Sie keine weiteren Fragen haben, dann sagen Sie bitte dem Versuchsleiter Bescheid und drücken anschließend auf Leertaste, damit die Übungsaufgaben beginnen können.

Instruction 3

Das war der Übungsdurchgang.

Wenn Sie noch Fragen haben, dann wenden Sie sich jetzt bitte an den Versuchsleiter.

Drücken Sie die Leertaste, um das Hauptexperiment zu starten.

Instruction 4

Vielen Dank für Ihre Teilnahme!

Bitte melden Sie sich beim Versuchsleiter.

Appendix A2: Stimuli of Experiment 1

Appendix A2

Stimuli of Experiment 1

Inference	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MP	N	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man lernt viel.	-	Schreibt man eine gute Klausur?
MP	S	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man lernt viel.	Die Klausur könnte schwierig sein.	Schreibt man eine gute Klausur?
MP	I	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man lernt viel.	Die Klausur ist schwierig.	Schreibt man eine gute Klausur?
MP	N	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Man düngt eine Pflanze.	-	Wächst die Pflanze schnell?
MP	S	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Man düngt eine Pflanze.	Die Pflanze könnte nicht genug Wasser bekommen.	Wächst die Pflanze schnell?
MP	I	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Man düngt eine Pflanze.	Die Pflanze bekommt nicht genug Wasser.	Wächst die Pflanze schnell?
MP	N	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel ist reif.	-	Fällt der Apfel vom Baum?
MP	S	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel ist reif.	Man könnte den Apfel pflücken.	Fällt der Apfel vom Baum?

Inference	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MP	I	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel ist reif.	Man pflückt den Apfel.	Fällt der Apfel vom Baum?
MP	N	Wenn man früh aufsteht, dann ist man müde.	Man steht früh auf.	-	Ist man müde?
MP	S	Wenn man früh aufsteht, dann ist man müde.	Man steht früh auf.	Man könnte Kaffee trinken.	Ist man müde?
MP	I	Wenn man früh aufsteht, dann ist man müde.	Man steht früh auf.	Man trinkt Kaffee.	Ist man müde?
MP	N	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Man wäscht Wäsche.	-	Wird die Wäsche sauber?
MP	S	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Man wäscht Wäsche.	Man könnte kein Waschmittel haben.	Wird die Wäsche sauber?
MP	I	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Man wäscht Wäsche.	Man hat kein Waschmittel.	Wird die Wäsche sauber?
MP	N	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Man streicht ein Streichholz an.	-	Gibt das Streichholz Feuer?
MP	S	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Man streicht ein Streichholz an.	Das Streichholz könnte nass sein.	Gibt das Streichholz Feuer?

Inference	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MP	I	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Man streicht ein Streichholz an.	Das Streichholz ist nass.	Gibt das Streichholz Feuer?
MT	N	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man schreibt keine gute Klausur.	-	Hat man viel gelernt?
MT	S	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man schreibt keine gute Klausur.	Die Klausur könnte schwierig gewesen sein.	Hat man viel gelernt?
MT	I	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man schreibt keine gute Klausur.	Die Klausur war schwierig.	Hat man viel gelernt?
MT	N	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Die Pflanze wächst nicht schnell.	-	Hat man die Pflanze gedüngt?
MT	S	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Die Pflanze wächst nicht schnell.	Die Pflanze könnte nicht genug Wasser bekommen haben.	Hat man die Pflanze gedüngt?
MT	I	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Die Pflanze wächst nicht schnell.	Die Pflanze hat nicht genug Wasser bekommen.	Hat man die Pflanze gedüngt?
MT	N	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel fällt nicht vom Baum.	-	Ist der Apfel reif gewesen?

Inference	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MT	S	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel fällt nicht vom Baum.	Der Apfel könnte gepflückt worden sein.	Ist der Apfel reif gewesen?
MT	I	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel fällt nicht vom Baum.	Der Apfel wurde gepflückt.	Ist der Apfel reif gewesen?
MT	N	Wenn man früh aufsteht, dann ist man müde.	Man ist nicht müde.	-	Ist man früh aufgestanden?
MT	S	Wenn man früh aufsteht, dann ist man müde.	Man ist nicht müde.	Man könnte Kaffee getrunken haben.	Ist man früh aufgestanden?
MT	I	Wenn man früh aufsteht, dann ist man müde.	Man ist nicht müde.	Man trank Kaffee.	Ist man früh aufgestanden?
MT	N	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Die Wäsche wird nicht sauber.	-	Hat man die Wäsche gewaschen?
MT	S	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Die Wäsche wird nicht sauber.	Man könnte kein Waschmittel gehabt haben.	Hat man die Wäsche gewaschen?
MT	I	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Die Wäsche wird nicht sauber.	Man hat kein Waschmittel gehabt.	Hat man die Wäsche gewaschen?
MT	N	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Das Streichholz gibt nicht Feuer.	-	Hat man das Streichholz angestrichen?

Inference	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MT	S	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Das Streichholz gibt nicht Feuer.	Das Streichholz könnte nass gewesen sein.	Hat man das Streichholz angestrichen?
MT	I	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Das Streichholz gibt nicht Feuer.	Das Streichholz war nass.	Hat man das Streichholz angestrichen?

Note. Inference: MP = Modus Ponens; MT = Modus Tollens. Mode: N = None; S = Subjunctive; I = Indicative.

Appendix A3: Response format of Experiment 1

Response format

Response format (counterbalanced)

Appendix B1: Instructions of Experiment 2

Instruction 1

Cari partecipanti,

qui di seguito vi vengono presentati dei compiti composti da diverse affermazioni. Le prime affermazioni sono scritte in nero e contengono un'affermazione del tipo „se…allora", un dato di fatto e, se necessario, ulteriori informazioni. L'ultima affermazione è scritta in rosso ed è una domanda che richiede una conclusione.

Il vostro compito è quello di rispondere a questa domanda. Avete a disposizione risposte da „no, in nessun caso" a „sì, in ogni caso" che vengono selezionate utilizzando il tastierino numerico. Il numero 1 sta per „no, in nessun caso" e il numero 7 per „sì, in ogni caso".

Si prega di rispondere come si farebbe in situazioni quotidiane.

Instruction 1 (counterbalanced)

Cari partecipanti,

qui di seguito vi vengono presentati dei compiti composti da diverse affermazioni. Le prime affermazioni sono scritte in nero e contengono un'affermazione del tipo „se…allora", un dato di fatto e, se necessario, ulteriori informazioni. L'ultima affermazione è scritta in rosso ed è una domanda che richiede una conclusione.

Il vostro compito è quello di rispondere a questa domanda. Avete a disposizione risposte da „sì, in ogni caso" a „no, in nessun caso" che vengono selezionate utilizzando il tastierino numerico. Il numero 1 sta per „sì, in ogni caso" e il numero 7 per „no, in nessun caso".

Si prega di rispondere come si farebbe in situazioni quotidiane.

Instruction 2

Per passare da un’affermazione all’altra si utilizza la barra spaziatrice e tra un compito e il successivo sarà data l’opportunità di fare una pausa. La pausa viene anche terminate utilizzando la barra spaziatrice. Si prega di leggere ogni compito con attenzione, perché le affermazioni possono anche contenere delle negazioni.

Tutto chiaro? Se non avete ulteriori domande, informate lo sperimentatore e poi premete la barra spaziatrice per iniziare il test di prova.

Instruction 3

Questo era il test di prova.

Se avete ancora domande, si prega di contattare adesso lo sperimentatore.

Premete la barra spaziatrice per avviare l’esperimento.

Instruction 4

Grazie della vostra partecipazione!

Si prega di contattare lo sperimentatore.

Appendix B2: Stimuli of Experiment 2

Appendix B2

Stimuli of Experiment 2

Inference	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MP	N	Se si studia tanto, si prende un buon voto all‘esame.	Si studia tanto.	-	Si prende un buon voto all‘esame?
MP	S	Se si studia tanto, si prende un buon voto all‘esame.	Si studia tanto.	L’esame potrebbe essere difficile.	Si prende un buon voto all‘esame?
MP	I	Se si studia tanto, si prende un buon voto all‘esame.	Si studia tanto.	L’esame è difficile.	Si prende un buon voto all‘esame?
MP	N	Se si concima una pianta, la pianta cresce velocemente.	Si concima una pianta.	-	La pianta cresce velocemente?
MP	S	Se si concima una pianta, la pianta cresce velocemente.	Si concima una pianta.	La pianta potrebbe non ottenere abbastanza acqua.	La pianta cresce velocemente?
MP	I	Se si concima una pianta, la pianta cresce velocemente.	Si concima una pianta.	La pianta non ottiene abbastanza acqua.	La pianta cresce velocemente?
MP	N	Se la mela è matura, la mela cade dall’albero.	La mela è matura.	-	La mela cade dall‘albero?
MP	S	Se la mela è matura, la mela cade dall’albero.	La mela è matura.	Si potrebbe raccogliere la mela.	La mela cade dall‘albero?
MP	I	Se la mela è matura, la mela cade dall’albero.	La mela è matura.	Si raccoglie la mela.	La mela cade dall‘albero?

Inference	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MP	N	Se ci si alza presto, si è stanchi.	Ci si alza presto.	-	Si è stanchi?
MP	S	Se ci si alza presto, si è stanchi.	Ci si alza presto.	Si potrebbe bere un caffè.	Si è stanchi?
MP	I	Se ci si alza presto, si è stanchi.	Ci si alza presto.	Si beve un caffè.	Si è stanchi?
MP	N	Se si lavano i panni, la biancheria si pulisce.	Si lavano i panni.	-	La biancheria si pulisce?
MP	S	Se si lavano i panni, la biancheria si pulisce.	Si lavano i panni.	Si potrebbero non avere detersivi.	La biancheria si pulisce?
MP	I	Se si lavano i panni, la biancheria si pulisce.	Si lavano i panni.	Non si hanno detersivi.	La biancheria si pulisce?
MP	N	Se si accende un fiammifero, il fiammifero dà fuoco.	Si accende un fiammifero.	-	Il fiammifero dà fuoco?
MP	S	Se si accende un fiammifero, il fiammifero dà fuoco.	Si accende un fiammifero.	Il fiammifero potrebbe essere bagnato.	Il fiammifero dà fuoco?
MP	I	Se si accende un fiammifero, il fiammifero dà fuoco.	Si accende un fiammifero.	Il fiammifero è bagnato.	Il fiammifero dà fuoco?
MT	N	Se si studia tanto, si prende un buon voto all’esame.	Non si prende un buon voto all’esame.	-	Si è studiato tanto?

Inference	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MT	S	Se si studia tanto, si prende un buon voto all'esame.	Non si prende un buon voto all'esame.	L'esame potrebbe essere stato difficile.	Si è studiato tanto?
MT	I	Se si studia tanto, si prende un buon voto all'esame.	Non si prende un buon voto all'esame.	L'esame era difficile.	Si è studiato tanto?
MT	N	Se si concima una pianta, la pianta cresce velocemente.	La pianta non cresce velocemente.	-	Si è concimata la pianta?
MT	S	Se si concima una pianta, la pianta cresce velocemente.	La pianta non cresce velocemente.	La pianta potrebbe non aver ottenuto abbastanza acqua.	Si è concimata la pianta?
MT	I	Se si concima una pianta, la pianta cresce velocemente.	La pianta non cresce velocemente.	La pianta non ha ottenuto abbastanza acqua.	Si è concimata la pianta?
MT	N	Se la mela è matura, la mela cade dall'albero.	La mela non cade dall'albero.	-	La mela era matura?
MT	S	Se la mela è matura, la mela cade dall'albero.	La mela non cade dall'albero.	La mela potrebbe essere stata raccolta.	La mela era matura?
MT	I	Se la mela è matura, la mela cade dall'albero.	La mela non cade dall'albero.	La mela è stata raccolta.	La mela era matura?
MT	N	Se ci si alza presto, si è stanchi.	Non si è stanchi.	-	Ci si è alzati presto?

Inference	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MT	S	Se ci si alza presto, si è stanchi.	Non si è stanchi.	Si potrebbe aver bevuto un caffè.	Ci si è alzati presto?
MT	I	Se ci si alza presto, si è stanchi.	Non si è stanchi.	Si è bevuto un caffè.	Ci si è alzati presto?
MT	N	Se si lavano i panni, la biancheria si pulisce.	La biancheria non si pulisce.	-	Si sono lavati i panni?
MT	S	Se si lavano i panni, la biancheria si pulisce.	La biancheria non si pulisce.	Si potrebbero non aver avuto detersivi.	Si sono lavati i panni?
MT	I	Se si lavano i panni, la biancheria si pulisce.	La biancheria non si pulisce.	Non si avevano detersivi.	Si sono lavati i panni?
MT	N	Se si accende un fiammifero, il fiammifero dà fuoco.	Il fiammifero non dà fuoco.	-	Si è acceso il fiammifero?
MT	S	Se si accende un fiammifero, il fiammifero dà fuoco.	Il fiammifero non dà fuoco.	Il fiammifero potrebbe essere stato bagnato.	Si è acceso il fiammifero?
MT	I	Se si accende un fiammifero, il fiammifero dà fuoco.	Il fiammifero non dà fuoco.	Il fiammifero era bagnato.	Si è acceso il fiammifero?

Note. Inference: MP = Modus Ponens; MT = Modus Tollens. Mode: N = None; S = Subjunctive; I = Indicative.

Appendix B3: Response format of Experiment 2

Response format

Response format (counterbalanced)

Appendix C1: Instructions of Experiment 3

Instruction 1

Lieber Teilnehmer, liebe Teilnehmerin,

im Folgenden werden Ihnen Aufgaben präsentiert, die sich aus mehreren Aussagen zusammensetzen. Die ersten Aussagen sind in schwarzer Schrift geschrieben und beinhalten eine Wenn-dann-Aussage, einen Fakt und gegebenenfalls eine Zusatzinformation. Die letzte Aussage ist in rot geschrieben und ist eine Frage, die nach einer Schlussfolgerung fragt.

Ihre Aufgabe besteht darin, diese Frage zu beantworten. Dazu stehen Ihnen Antworten von „nein, auf keinen Fall“ bis „ja, auf jeden Fall“ zur Verfügung, die Sie mit Hilfe der grünen Ziffernreihe auf der Tastatur angeben. Die Nummer 1 steht hierbei für „nein, auf keinen Fall“ und die Nummer 7 für „ja, auf jeden Fall“.

Antworten Sie bitte so, wie Sie es auch in Alltagssituationen machen würden.

Instruction 1 (counterbalanced)

Lieber Teilnehmer, liebe Teilnehmerin,

im Folgenden werden Ihnen Aufgaben präsentiert, die sich aus mehreren Aussagen zusammensetzen. Die ersten Aussagen sind in schwarzer Schrift geschrieben und beinhalten eine Wenn-dann-Aussage, einen Fakt und gegebenenfalls eine Zusatzinformation. Die letzte Aussage ist in rot geschrieben und ist eine Frage, die nach einer Schlussfolgerung fragt.

Ihre Aufgabe besteht darin, diese Frage zu beantworten. Dazu stehen Ihnen Antworten von „ja, auf jeden Fall“ bis „nein, auf keinen Fall“ zur Verfügung, die Sie mit Hilfe der grünen Ziffernreihe auf der Tastatur angeben. Die Nummer 1 steht hierbei für „ja, auf jeden Fall“ und die Nummer 7 für „nein, auf keinen Fall“.

Antworten Sie bitte so, wie Sie es auch in Alltagssituationen machen würden.

Instruction 2

Von Aussage zu Aussage kommen Sie immer mit der Leertaste weiter. Zwischen den Aufgaben erscheint ein Fixationspunkt. Hier können Sie bei Bedarf eine Pause machen. Die Pause beenden Sie ebenfalls mit der Leertaste. Lesen Sie bitte jede Aufgabe sorgfältig durch, da die Aussagen auch Verneinungen enthalten können.

Alles klar? Wenn Sie keine weiteren Fragen haben, dann sagen Sie bitte dem Versuchsleiter Bescheid und drücken anschließend auf Leertaste, damit die Übungsaufgaben beginnen können.

Instruction 3

Das war der Übungsdurchgang.

Wenn Sie noch Fragen haben, dann wenden Sie sich jetzt bitte an den Versuchsleiter.

Drücken Sie die Leertaste, um das Hauptexperiment zu starten.

Instruction 4

Der Experimentalblock ist beendet.

Abschließend bitten wir Sie auf der nachfolgenden Seite um demografische Angaben.

Bitte machen Sie die Angaben vollständig!

Instruction 5

Vielen Dank für Ihre Teilnahme!

Bitte melden Sie sich beim Versuchsleiter.

Appendix C2: Stimuli of Experiment 3

Appendix C2

Stimuli of Experiment 3

Inference	Relevance	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MP	N	N	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man lernt viel.	-	Schreibt man eine gute Klausur?
MP	I	S	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man lernt viel.	Der Ball könnte grün sein.	Schreibt man eine gute Klausur?
MP	I	I	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man lernt viel.	Der Ball ist grün.	Schreibt man eine gute Klausur?
MP	R	S	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man lernt viel.	Die Klausur könnte schwierig sein.	Schreibt man eine gute Klausur?
MP	R	I	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man lernt viel.	Die Klausur ist schwierig.	Schreibt man eine gute Klausur?
MP	N	N	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Man düngt eine Pflanze.	-	Wächst die Pflanze schnell?
MP	I	S	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Man düngt eine Pflanze.	Der neue Laptop könnte glänzend grau sein.	Wächst die Pflanze schnell?

Inference	Relevance	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MP	I	I	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Man düngt eine Pflanze.	Der neue Laptop ist glänzend grau.	Wächst die Pflanze schnell?
MP	R	S	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Man düngt eine Pflanze.	Die Pflanze könnte nicht genug Wasser bekommen.	Wächst die Pflanze schnell?
MP	R	I	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Man düngt eine Pflanze.	Die Pflanze bekommt nicht genug Wasser.	Wächst die Pflanze schnell?
MP	N	N	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel ist reif.	-	Fällt der Apfel vom Baum?
MP	I	S	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel ist reif.	Das Telefon könnte schwarz sein.	Fällt der Apfel vom Baum?
MP	I	I	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel ist reif.	Das Telefon ist schwarz.	Fällt der Apfel vom Baum?
MP	R	S	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel ist reif.	Man könnte den Apfel pflücken.	Fällt der Apfel vom Baum?
MP	R	I	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel ist reif.	Man pflückt den Apfel.	Fällt der Apfel vom Baum?

Inference	Relevance	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MP	N	N	Wenn man früh aufsteht, dann ist man müde.	Man steht früh auf.	-	Ist man müde?
MP	I	S	Wenn man früh aufsteht, dann ist man müde.	Man steht früh auf.	Die Schaufel könnte beige sein.	Ist man müde?
MP	I	I	Wenn man früh aufsteht, dann ist man müde.	Man steht früh auf.	Die Schaufel ist beige.	Ist man müde?
MP	R	S	Wenn man früh aufsteht, dann ist man müde.	Man steht früh auf.	Man könnte Kaffee trinken.	Ist man müde?
MP	R	I	Wenn man früh aufsteht, dann ist man müde.	Man steht früh auf.	Man trinkt Kaffee.	Ist man müde?
MP	N	N	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Man wäscht Wäsche.	-	Wird die Wäsche sauber?
MP	I	S	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Man wäscht Wäsche.	Das Auto könnte blau sein.	Wird die Wäsche sauber?
MP	I	I	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Man wäscht Wäsche.	Das Auto ist blau.	Wird die Wäsche sauber?
MP	R	S	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Man wäscht Wäsche.	Man könnte kein Waschmittel haben.	Wird die Wäsche sauber?

Inference	Relevance	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MP	R	I	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Man wäscht Wäsche.	Man hat kein Waschmittel	Wird die Wäsche sauber?
MP	N	N	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Man streicht ein Streichholz an.	-	Gibt das Streichholz Feuer?
MP	I	S	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Man streicht ein Streichholz an.	Die Tasse könnte lila sein.	Gibt das Streichholz Feuer?
MP	I	I	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Man streicht ein Streichholz an.	Die Tasse ist lila.	Gibt das Streichholz Feuer?
MP	R	S	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Man streicht ein Streichholz an.	Das Streichholz könnte nass sein.	Gibt das Streichholz Feuer?
MP	R	I	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Man streicht ein Streichholz an.	Das Streichholz ist nass.	Gibt das Streichholz Feuer?
MT	N	N	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man schreibt keine gute Klausur.	-	Hat man viel gelernt?
MT	I	S	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man schreibt keine gute Klausur.	Der Ball könnte grün gewesen sein.	Hat man viel gelernt?

Inference	Relevance	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MT	I	I	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man schreibt keine gute Klausur.	Der Ball war grün.	Hat man viel gelernt?
MT	R	S	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man schreibt keine gute Klausur.	Die Klausur könnte schwierig gewesen sein.	Hat man viel gelernt?
MT	R	I	Wenn man viel lernt, dann schreibt man eine gute Klausur.	Man schreibt keine gute Klausur.	Die Klausur war schwierig.	Hat man viel gelernt?
MT	N	N	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Die Pflanze wächst nicht schnell.	-	Hat man die Pflanze gedüngt?
MT	I	S	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Die Pflanze wächst nicht schnell.	Der neue Laptop könnte glänzend grau gewesen sein.	Hat man die Pflanze gedüngt?
MT	I	I	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Die Pflanze wächst nicht schnell.	Der neue Laptop war glänzend grau.	Hat man die Pflanze gedüngt?
MT	R	S	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Die Pflanze wächst nicht schnell.	Die Pflanze könnte nicht genug Wasser bekommen haben.	Hat man die Pflanze gedüngt?
MT	R	I	Wenn man eine Pflanze düngt, dann wächst die Pflanze schnell.	Die Pflanze wächst nicht schnell.	Die Pflanze hat nicht genug Wasser bekommen.	Hat man die Pflanze gedüngt?

Inference	Relevance	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MT	N	N	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel fällt nicht vom Baum.	-	Ist der Apfel reif gewesen?
MT	I	S	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel fällt nicht vom Baum.	Das Telefon könnte schwarz gewesen sein.	Ist der Apfel reif gewesen?
MT	I	I	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel fällt nicht vom Baum.	Das Telefon war schwarz.	Ist der Apfel reif gewesen?
MT	R	S	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel fällt nicht vom Baum.	Der Apfel könnte gepflückt worden sein.	Ist der Apfel reif gewesen?
MT	R	I	Wenn der Apfel reif ist, dann fällt der Apfel vom Baum.	Der Apfel fällt nicht vom Baum.	Der Apfel wurde gepflückt.	Ist der Apfel reif gewesen?
MT	N	N	Wenn man früh aufsteht, dann ist man müde.	Man ist nicht müde.	-	Ist man früh aufgestanden?
MT	I	S	Wenn man früh aufsteht, dann ist man müde.	Man ist nicht müde.	Die Schaufel könnte beige gewesen sein.	Ist man früh aufgestanden?
MT	I	I	Wenn man früh aufsteht, dann ist man müde.	Man ist nicht müde.	Die Schaufel war beige.	Ist man früh aufgestanden?
MT	R	S	Wenn man früh aufsteht, dann ist man müde.	Man ist nicht müde.	Man könnte Kaffee getrunken haben.	Ist man früh aufgestanden?

Inference	Relevance	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MT	R	I	Wenn man früh aufsteht, dann ist man müde.	Man ist nicht müde.	Man trank Kaffee.	Ist man früh aufgestanden?
MT	N	N	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Die Wäsche wird nicht sauber.	-	Hat man Wäsche gewaschen?
MT	I	S	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Die Wäsche wird nicht sauber.	Das Auto könnte blau gewesen sein.	Hat man Wäsche gewaschen?
MT	I	I	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Die Wäsche wird nicht sauber.	Das Auto war blau.	Hat man Wäsche gewaschen?
MT	R	S	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Die Wäsche wird nicht sauber.	Man könnte kein Waschmittel gehabt haben.	Hat man Wäsche gewaschen?
MT	R	I	Wenn man Wäsche wäscht, dann wird die Wäsche sauber.	Die Wäsche wird nicht sauber.	Man hat kein Waschmittel gehabt.	Hat man Wäsche gewaschen?
MT	N	N	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Das Streichholz gibt kein Feuer.	-	Hat man das Streichholz angestrichen?
MT	I	S	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Das Streichholz gibt kein Feuer.	Die Tasse könnte lila gewesen sein.	Hat man das Streichholz angestrichen?

Inference	Relevance	Mode	Major Premise	Minor Premise	Counterargument	Conclusion
MT	I	I	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Das Streichholz gibt kein Feuer.	Die Tasse war lila.	Hat man das Streichholz angestrichen?
MT	R	S	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Das Streichholz gibt kein Feuer.	Das Streichholz könnte nass gewesen sein.	Hat man das Streichholz angestrichen?
MT	R	I	Wenn man ein Streichholz anstreicht, dann gibt das Streichholz Feuer.	Das Streichholz gibt kein Feuer.	Das Streichholz war nass.	Hat man das Streichholz angestrichen?

Note. Inference: MP = Modus Ponens; MT = Modus Tollens. Relevance: N = None; I = Irrelevant; R = Relevant. Mode: N = None; S = Subjunctive; I = Indicative.

Appendix C3: Response format of Experiment 3

Response format

Response format (counterbalanced)

Declaration of Originality

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel *Aspects of Rational Argumentation* selbstständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen 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 Angabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Tübingen, 01.06.2022

Bruno Richter