Information-Theoretic Perspectives on Unconscious Priming and Group Decisions: Retrieving Maximum Information From Human Responses

Dissertation

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

vorgelegt von
Sascha Meyen
aus Potsdam

Tübingen
2021

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

Tag der mündlichen Qualifikation	20.12.2021
Dekan:	Prof. Dr. Thilo Stehle
1. Betreuer:	Prof. Dr. Volker H. Franz
2. Betreuerin:	Prof. Dr. Ulrike von Luxburg

Abstract

Since Information Theory was introduced to the field of psychology in the middle of the last century, Information processing in the human brain has gained attention. A question of particular interest has been: To what degree can humans process information unconsciously? For the past two decades, one of the most prominent paradigms in which this question has been investigated was *unconscious priming*. Studies in this paradigm have frequently used a *standard reasoning*: These studies show that participants perform close to random guessing when they have to identify barely visible stimuli but the same stimuli nevertheless produce clear effects in indirect measures such as reaction times or neuroimaging measures. From this pattern of results, the standard reasoning concludes that participants processed information about the stimuli beyond what they are consciously aware of. But we show here that the standard reasoning is flawed. The clear effects in indirect measures can often be fully explained by residual conscious processing that is reflected in participants' close to (but not exactly equal to) chance level guessing in consciously given responses. The erroneous appearance of more unconscious processing is due to an inappropriate comparison making conscious responses appear as if they were based on less information than the indirect measures. We develop a reanalysis method for results from these studies and demonstrate that a large body of heavily cited literature in the paradigm has little to no evidence for their strong claims on unconscious processing.

In the field of group decision making, a similar methodological problem occurs. Here, researchers aim to model real group discussions via statistical aggregations of individual group members' responses. The statistically aggregated responses serve as *simulated* group decisions that are then compared to the *real* group decisions coming from an interactive group discussion. A common result is that real group decisions are more accurate than simulated group decisions. But most studies do not use the theoretically optimal method of Confidence Weighted Majority Voting (CWMV) to simulate group decisions. Similar to unconscious priming, suboptimal methods for simulations lead to inappropriate comparisons between simulated vs. real decisions. This

in turn may lead to unwarranted interpretations due to methodological bias. We bring forward the theoretically optimal method and demonstrate in an experiment that simulated and real group decisions are equally accurate with this method. Despite matching in accuracy, real groups systematically deviate from CWMV simulations. We capture these deviations in a formal cognitive model showing that real groups treat each group member's vote more equally than CWMV predicts. Moreover, real groups exhibit an overall lower confidence than CWMV simulations. What ties group decision making and unconscious priming together is that the full information from human participants' responses was not fully taken into account when making comparisons.

Our results raise the additional question of whether, based on the accuracies of the individuals, we can a priori determine the *accuracy of the group*. This is particularly interesting in machine learning where not individual humans form a group but individual classifiers form an ensemble. We introduce a model in which we demonstrate a negative result: The ensemble accuracy can take values in a surprisingly large range even when the individual classifiers' accuracies are held constant. This is because individual classifiers with a fixed accuracy can still convey drastically varying amounts of information. We prove best- and worst-case ensemble accuracies for when the individual classifiers' accuracies are known. Additionally, we provide tighter bounds for cases in which not only accuracies but also individual classifiers' transmitted information is known. Our constructive proofs yield guiding principles for selecting and constructing classifiers for ensembles. These principles go beyond the simple notion of preferring classifiers with highest mutual information.

These three strands of research highlight the relevance of certain aspects from responses given by humans or classifiers that go beyond classification accuracy. Such aspects are prima facie easily overlooked in many scenarios. But they still affect mutual information measures and can have theoretical and practical impact as we demonstrate in unconscious priming, decision making, and ensemble accuracy.

Zusammenfassung

Seitdem die Informationstheorie in die psychologische Forschungsliteratur während der Mitte des letzten Jahrhunderts eingeführt worden ist, hat das Thema Informationsverarbeitung im menschlichen Gehirn mehr und mehr an Aufmerksamkeit gewonnen. Von besonderem Interesse war die Frage, zu welchem Ausmaß Menschen Informationen unterbewusst verarbeiten können. Um unterbewusste Informationsverarbeitung nachzuweisen, war in den letzten zwei Jahrzehnten eines der prominentesten Paradigmen das unterbewusste Priming. Studien in diesem Paradigma folgen häufig einem *Standardverfahren*: Diese Studien zeigen, dass Versuchsteilnehmer nahe am Rateniveau sind, wenn sie kaum sichtbare Stimuli identifizieren sollen. Gleichzeitig produzieren dieselben Stimuli eindeutige Effekte in indirekten Messungen wie zum Beispiel in Reaktionszeiten oder in Gehirnaktivierung. Von diesem Ergebnismuster wird im Standardverfahren geschlussfolgert, dass die Versuchsteilnehmer mehr Information über die Stimuli verarbeitet haben, als ihnen bewusst ist. Wir zeigen hier, dass das Standardverfahren fehlerhaft ist. Die Effekte auf die indirekten Messungen können meist vollständig durch residuale, bewusste Verarbeitung erklärt werden. Diese schwache, bewusste Verarbeitung zeigt sich in der Identifikationsleistung der Studienteilnehmer, welche zwar nahe am Zufallsniveau aber doch nicht exakt auf diesem liegt. Der irreführende Eindruck einer überlegenen, unterbewussten Verarbeitung entsteht durch einen methodisch unangemessenen Vergleich. Dabei erscheinen die direkt gegebenen Antworten, als basierten sie auf weniger Information über die Stimuli als die indirekten Messungen. Wir entwickeln hier eine Reanalysemethode für Ergebnisse aus früheren Studien und zeigen, dass große Teile der vielzitierten Forschungsliteratur zu unterbewusstem Priming wenig bis keine Beweise für die weitreichenden Interpretationen über unterbewusste Verarbeitung liefern.

Im Forschungsfeld Gruppenentscheidungen gibt es einen analogen Fehler. In solchen Studien werden echte Gruppenentscheidungsprozesse simuliert, indem Aussagen individueller Gruppenmitglieder statistisch zusammengeführt werden. Diese statistischen Zusammenführungen dienen als *simulierte* Gruppenentscheidungen, die dann mit den *echten* Gruppenentscheidungen aus inter-

aktiven Gruppendiskussionen verglichen werden. Ein Ergebnis solcher Studien ist, dass echte Gruppen häufiger korrekte Entscheidungen treffen als simulierte Gruppen. Aber die meisten Studien nutzen nicht die theoretisch optimale Methode, Mehrheitsbeschluss mit Stimmgewichtung (Confidence Weighted Majority Voting, CWMV). Ähnlich zum Problem beim unterbewussten Priming führen suboptimale Methoden bei der Simulation von Gruppenentscheidungen potenziell zu ungerechtfertigten Interpretationen beim Vergleich von echten mit simulierten Gruppen. Unterschiede können allein durch methodische Disparitäten auftreten und dürfen nicht ohne weiteres auf einen zugrundeliegenden, wahren Unterschied zwischen echten und simulierten Gruppen zurückgeführt werden. Wir stellen die theoretisch optimale Methode CWMV in den Blickpunkt und zeigen in einem Experiment, dass diese Methode gleiche Vorhersagegenauigkeit wie echte Gruppenentscheidungen erreicht. Das macht CWMV zu einem geeigneten Kandidaten, um echte Gruppenprozesse zu modellieren. Obwohl die Vorhersagegenauigkeit übereinstimmt, unterscheiden sich echte Gruppen systematisch von den Simulationen mit CWMV. Wir modellieren diese Abweichungen und zeigen, dass echte Gruppen die Aussagen ihrer Mitglieder gleichmäßiger gewichten und insgesamt weniger Sicherheitsbewertung in die Gruppenentscheidung legen als Simulationen mittels CWMV. Beide Forschungsbereiche – unterbewusstes Priming und Gruppenentscheidungsprozesse – vereint, dass die volle Information in den Antworten der Versuchsteilnehmer oft nicht vollständig berücksichtigt wird.

Unsere Ergebnisse im Bereich der Gruppenentscheidungen werfen die zusätzliche Frage auf, ob die Vorhersagegenauigkeit der einzelnen Gruppenmitglieder die *Vorhersagegenauigkeit der Gruppe* bestimmt. Diese Frage ist insbesondere im Bereich des maschinellen Lernens relevant, bei der nicht Menschen eine Gruppe bilden sondern einzelne Klassifikationsalgorithmen ein sogenanntes Ensemble. Wir erarbeiten hier ein Modell, in dem wir ein Negativergebnis nachweisen: Die Vorhersagegenauigkeit des Ensembles kann Werte in einer überraschend breiten Spanne annehmen, selbst wenn die Vorhersagegenauigkeit der einzelnen Klassifikationsalgorithmen konstant gehalten wird. Der Grund liegt in dem drastisch unterschiedlichen Informationsgehalt, den ein Klassifikationsalgorithmus trotz gleich gehaltener Genauigkeit übertragen kann. Wir beweisen, welche Vorhersagegenauigkeit eine Ensemble im besten und schlechtesten Fall bei gegebener Genauigkeit der einzelnen Algorithmen annehmen kann. Zusätzlich beweisen wir engere Schranken für den Fall, dass nicht nur die Klassifikationsgenauigkeit sondern auch der Informationsgehalt der einzelnen Algorithmen gegeben ist. Aus unseren konstruktiven Beweisen gehen Prinzipien für die Auswahl und Implementation von Klassifikationsalgorithmen für die Verwendung in Ensembles hervor. Diese Prinzipien gehen über

die einfache Heuristik hinaus, dass Klassifikationsalgorithmen mit höherem Informationsgehalt gewählt werden sollten.

Diese drei Forschungsgegenstände unterstreichen die Relevanz von Aspekten menschlicher und maschineller Vorhersagen, welche jenseits der Vorhersagegenauigkeit liegen. Diese Aspekte sind auf den ersten Blick leicht übersehen, da viele psychologische Forschungsbereiche sich auf das herkömmliche Maß der Klassifikationsgenaugkeit beschränken. Sie spielen nichtsdestotrotz eine wichtige thereotische und praktische Rolle, wie wir in den drei Bereichen zeigen.

Contents

Chapter 1

Introduction

Research objects are inseparable from the research methods used to study them. Using one particular research method only allows for a limited perspective on the research objects. Thus, new methods often not only improve precision but enrich those perspectives. For example, the development of the microscope has not only enhanced precision but opened a new view on microbiology. In the same way, advances in methods for psychological research transform how researchers understand psychological processes.

We analyze methods in two fields of psychological research. First in *unconscious priming* and second in *group decision making*. What ties these two fields together is that methods from both can be better understood by looking at them through the lens of *Information Theory* (Shannon, 1948; Cover & Thomas, 2006; MacKay, 2003). We borrow the measure of mutual information from this theory and use it to assess how much information human participants can report. This way, we reveal that participants can report information that has not been properly accounted for by current methods. We present methods that can take this additional information into account and discuss implications for psychological research.

1.1 Information-Theoretic Measures for Human Responses

Information Theory with its associated measure of mutual (or transmitted) information has previously been applied in other subfields of psychology (Attneave, 1959; Garner, 1962). Originally, Shannon had measured redundancy of natural language utterances (Shannon, 1948). In principle, each letter in the English language can transmit $\log_2(26) = 4.7$ bit of information because choosing one out of 26 letters is similar to answering an average of 4.7 binary questions.

But Shannon found that natural languages are surprisingly redundant. For example, the string "redundan—" from the previous sentence must have been followed by "t" making the last letter redundant; also "t" is one of the most frequent letters further increasing its redundancy. Shannon's result was that each letter transmits much less information than it could. This should not be seen as a flaw in our use of language but a way to ensure stable transmission of information through language. But independent of its interpretation, the idea of measuring how much information humans transmit through their deliberate responses has propelled research.

Soon after Shannon's seminal introduction, the mutual information measure was used to measure the correspondence between stimuli presented to human participants and their response to these stimuli. For example, Hake and Garner (1951) investigated how many different buttons participants can effectively use (see also Garner, 1960). The study had participants identify stimuli by pressing one of multiple buttons. In theory, by giving more buttons more information can be transmitted (much like increasing the number of letters in an alphabet). But practically, more buttons also prompt participants to make more mistakes in identifying the stimuli. The study's striking result was that, when increasing the number of buttons available to the participants, the mutual information plateaued at around 3.2 bit. This corresponds to participants effectively using only around nine buttons. Similar research found a transmitted information rate of 2.2 bit (five buttons) for auditory stimuli (Pollack, 1952; Eriksen & Hake, 1955a). Much higher rates of transmitted information can be found for multidimensional stimuli (Klemmer & Frick, 1953; Pollack & Ficks, 1954; Eriksen & Hake, 1955b).

Information is not only related to participants' deliberate responses but also to response times. Two prominent psychophysical laws describe these relationships. First, when participants have to identify one stimulus out of a number of possible alternatives, Hick's Law relates their reaction times to the processed information (Hick, 1952; Hyman, 1953). In particular, reaction times scale linearly with the information provided by the participants' responses (1 bit for two alternatives, 2 bit for four alternatives, 3 bit for eight alternatives, etc.). In Section 2.6.4, we will encounter Hick's paradigm for measuring reaction times again. Second, when participants have to point to continuously variable stimulus locations, Fitt's Law relates reaction times to the continuous stimuli alternatives (Fitts, 1954; Shannon, 1949). Again, this law describes a linear relationship between reaction times and the information provided by discriminating target locations of variable sizes.

Since the rise of neurosciences, the mutual information measure has been primarily used as a principled metric to study processes in the human brain (see

Piasini & Panzeri, 2019, and the associated Special Issue; Bonnasse-Gahot & Nadal, 2008; Kang, Shapley, & Sompolinsky, 2004). The idea of measuring how much information processing in bits is, by now, ingrained in neuroscience. One of the most prominent theories of consciousness, the Integrated Information Theory, bases its definition of consciousness on the mutual information between brain areas and how they are causally interconnected (Tononi, 2004; Oizumi, Albantakis, & Tononi, 2014).

Nevertheless, behavioral studies continue to investigate the relation between stimulus information and human participants' responses (Bates, Lerch, Sims, & Jacobs, 2019). Our focus here will be more on human participants' behavioral responses: deliberate decisions and reaction times. Studying reaction times has a long history when it comes to inferring information processing in the human brain (Luce et al., 1986). But also neuroscience topics will be discussed in Chapter 2 and our results there apply to studies with neuroimaging results as well.

1.2 Comparative Approaches Through the Lens of Information Theory

In this thesis, we focus on methodological issues revolving around the information transmitted through participants' responses. In particular, we will look at methods in the two fields, unconscious priming and group decision making. Methods in both fields compare how much information human participants process in different tasks. In both fields, participants seem to be better in one task as compared with another suggesting underlying differences in human information processing between these two tasks. However, we will argue that some of the information human participants' can provide is overlooked in one task. This may lead to problematic conclusions because when information is neglected in one task, an apparent difference between the two tasks is no valid indicator for a true underlying difference.

In the first field, unconscious priming research, we scrutinize a prominent method used to demonstrate that processing occurred unconsciously (Hannula, Simons, & Cohen, 2005; Holender, 1986; Merikle, 1992). We make this method's internal logic explicit and reveal a fallacy that has led to unwarranted claims about unconscious processing in many domains (Dehaene et al., 1998; ten Brinke, Stimson, & Carney, 2014; Pessiglione et al., 2007; Wójcik, Nowicka, Bola, & Nowicka, 2019; van Gaal, Ridderinkhof, Scholte, & Lamme, 2010). At the core of this method is the comparison of how much information was processed in a conscious task vs. an unconscious task. But this comparison

is made improperly. The current method makes it seem as if more information was processed unconsciously than consciously—even if participants have full conscious access to all the processed information. We demonstrate with an example, how the apparently different results in both tasks can come to place even when the same underlying information generated those results. To overcome the problems of this method, we develop an appropriate comparison between the conscious and unconscious tasks. Our approach provides a way for future unconsciousness research to follow the often ignored requests to make this comparison proper (Eriksen, 1960; Merikle & Reingold, 1998; Reingold & Merikle, 1988). With our method, we investigate whether there really is evidence for information processing that exceeds what participants can consciously report and which would therefore be classified as unconscious. As a result, we find no empirical evidence for unconscious information processing in many studies contradicting their original claims. Our results challenge interpretations made by many studies from the past two decades that considered unconscious processing in some way superior to conscious processing such as the Unconscious Thought Theory (Dijksterhuis, Bos, Nordgren, & Van Baaren, 2006) or the Yes-It-Can(-be-processed-unconsciously) principle (Hassin, 2013) and others (ten Brinke, Vohs, & Carney, 2016).

In the second field of group decision making research, group decisions are obtained from two conditions and then compared. First, human participants give individual responses (decisions). These decisions are then statistically aggregated into a simulated group decision. This can be done for example by a simple majority vote, taking the most frequently reported decision as the simulated group decision (Ladha, 1992). Second, the same individual participants are then conducting a real group discussion and come up with a real group decision (e.g., Sniezek & Henry, 1990). A common result is that real groups outperform simulated groups in that their decisions are more often correct. This lends credibility to the idea of synergy that only occurs in real group discussions. But the problem with such interpretations is that information from human participants' reports is often discarded or not handled properly. Mainly, there are many situations in which human participants can estimate how reliable their individual responses are in the form of additional confidence ratings (Bahrami et al., 2010; Brenner, Koehler, Liberman, & Tversky, 1996; Griffin & Tversky, 1992; Zehetleitner & Rausch, 2013). In these situations, a majority vote would discard additional information from these confidence ratings because a majority vote weights each vote equally. We put forward the method *Confidence Weighted Majority Voting* (CWMV) which originated from mathematics (Grofman, Owen, & Feld, 1983; Nitzan & Paroush, 1982). This is the theoretically optimal method to combine reports of individual group

members when they include confidence ratings. Studies in the field sometimes do not use confidence ratings at all or use them in a suboptimal way (e.g., consider the methods in Hastie & Kameda, 2005; Hautz, Kämmer, Schauber, Spies, & Gaissmaier, 2015; van Dijk, Sonnemans, & Bauw, 2014; Kosinski, Bachrach, Kasneci, Van-Gael, & Graepel, 2012; Kurvers et al., 2016; Sorkin, Hays, & West, 2001). Then, it should come as no surprise that real groups outperform simulated groups because not all the information (including meta-information via confidences) enters the simulated group decisions properly. We implemented the CWMV simulation method in an experiment to test whether human participants are able to make optimal use of the information from individual confidence ratings. In our experiment, real group decisions were as accurate as simulated group decisions. We found no superiority of real groups over the simulated groups. However, we also found that real groups diverge to some degree from this optimal method. We use a cognitive model to adapt CWMV to the data so that we can better describe and simulate real group decisions.

Considering how the information of individual reports affects group decision making has directed us to a third research question. In many tasks, individuals are not as reliable as desired and are replaced by a group to increase accuracy (just consider the example in Hautz et al., 2015, where groups of medical students provide better diagnoses than they do individually). The question here is: Can we predict the accuracy of the group given that we know the accuracy of each individual member? This question is relevant beyond human groups and for machine learning where not human individuals convene as a group but instead individual classifiers form an ensemble (Hansen & Salamon, 1990; Lakshminarayanan, Pritzel, & Blundell, 2017; Dietterich, 2000; Zhou, 2012). This has already been answered for simple situations in which individuals do not vary in their competence from decision to decision (Ladha, 1992; Grofman et al., 1983). In these settings, the group accuracy can be exactly computed from the individual accuracies. However, we extend this setting to a more natural setting in which individuals may vary in their competence from decision to decision. There, we answer this question with a negative result: Even when the accuracies of the individuals are known, the ensemble accuracy is not uniquely determined but there is a surprisingly large range of possible values the ensemble accuracies can take. This is because individuals with the same accuracy may still differ in the amount of information they provide. When individuals can distinguish which of their responses are reliable and which are not, they provide more information than when they cannot make this distinction—even if they equally often respond correctly. As a consequence, ensembles of more informative individuals produce much better ensemble decisions than their less

informative counterparts. Even when the individuals have the same accuracy, ensembles of more informative individuals end up having higher accuracy. This is in line with previous research showing the benefit of feature selection based on mutual information (Battiti, 1994; Estévez, Tesmer, Perez, & Zurada, 2009; Vergara & Estévez, 2014; a single feature can be interpreted as an individual member and the whole feature set corresponds to the ensemble). We provide bounds for the best- and worst-case ensemble accuracies for when the accuracies of the individual are known. We obtain these bounds constructively by modeling how most informative and least informative individuals are constituted. These construction principles are relevant for training and selecting individual machine learning classifiers. Our construction principles go beyond the established result that more mutual information (given constant accuracy) is beneficial. In some cases, the best choice of an individual in a group can have *lower* accuracy and *equal* mutual information compared to other individuals as long as it follows our optimal construction principle.

The following three introductory sections summarize our contributions in the three research topics: (1) unconscious priming, (2) group decision making, and (3) ensemble accuracy bounds.

1.3 Unconscious Priming

Studies in the field of unconscious priming attempt to demonstrate that humans process information outside conscious awareness (Hannula et al., 2005; Simons, Hannula, Warren, & Day, 2007). They frequently use a method that we will call the *standard reasoning* of unconscious priming. It goes like this: Human participants have to complete two tasks. One task is meant to measure conscious processing and the other is meant to measure unconscious processing. The conscious task requires participants to discriminate barely visible stimuli directly. The unconscious task measures how well participants discriminate the stimuli in an indirect way. The typical conclusion is that indirect measures reveal that participants discriminate the presented stimuli better than they can report when asked directly. As Dell'Acqua and Grainger (1999) put it: "More specifically, null effects are sought in direct measures (i.e. where subjects respond directly to the unconsciously presented stimuli) accompanied by non-null indirect effects (i.e. priming effects)" (p. B2). This pattern of results lends credibility to the notion that participants unconsciously processed information that goes beyond what they are consciously aware of. For similar accounts of this standard reasoning, see Erdelyi (1986); Goldiamond (1958); Klotz and Neumann (1999); Lau (2007); Naccache and Dehaene (2001a); Peremen and Lamy (2014). Holender (1986) labeled this standard reasoning *se-*

mantic activation without conscious identification, where semantic activation is measured indirectly via reaction times or neuroimaging.

We focus on one example study of the paradigm: the seminal work by Dehaene et al. (1998). This study investigated unconscious processing of masked number stimuli. Following the standard reasoning, they presented number stimuli in two tasks. In the conscious task, which we will also refer to as the direct task, they presented a number stimulus from 1 to 9 briefly (for only 43 ms) and surrounded by mask stimuli such that it was hardly visible to the participants. When participants are asked to identify whether a stimulus was larger or smaller than 5, they performed close to chance level—not much better than if they were randomly guessing. In the unconscious task, participants were presented with the same masked stimulus but had to respond to a second, subsequently presented number stimulus. The second stimulus was presented long enough to be clearly visible. Even though participants had to respond to the second stimulus while the first being irrelevant for this task, their responses were indirectly affected by the first. In particular, reaction times were faster when both stimuli were semantically congruent (both larger or also smaller than 5) than when they were incongruent (one stimulus smaller and the other larger than 5). Hence the first stimulus had an indirect effect on reaction times by priming participant's responses. Because participants did not show above chance performance in the direct (conscious) task but a clear priming effect in the indirect (unconscious) task, the authors concluded that processing of the first number stimulus occurred unconsciously.

We will follow the tradition of unconscious priming studies and use Signal Detection Theory (Green & Swets, 1988; Macmillan & Creelman, 2004) to measure the sensitivity (d'). This is an indicator of how much information participants processed of the barely visible, masked stimuli. The methods used in the two tasks differ. This leads to an inherent difference in results for purely methodological reasons. One difference between the two tasks is that they measure different amounts of information. In the conscious task, participants are typically required to give binary responses when trying to identify the masked stimulus ("Is the stimulus from category A or B?"). Participants thus cannot give a continuous response, e.g., about how confident they are in their decisions ("I am very sure the stimulus is from category A" vs. "I am unsure, but if I have to guess, I'd say A"). This meta-information in form of confidence is receiving increasing attention in recent years (Mamassian, 2016; Rahnev et al., 2020; Zehetleitner & Rausch, 2013; but see also the early work of Peirce & Jastrow, 1884). In contrast, the unconscious task typically involves continuous measures such as reaction times. It is a priori known that continuous measures can carry more information than binary measures (Cohen, 1983;

Fedorov, Mannino, & Zhang, 2009). In our case, very fast (vs. slow) reaction times indicate congruent (vs. incongruent) trials with relatively high certainty while intermediate reaction times come with low certainty. By not asking participants for a continuous confidence measure in the direct task, this additional information of discriminating trials with high vs. low confidence is discarded. See Figure 1.1 for a quantification of the lost information—difference between solid and dashed lines—taken from our previous work (Meyen, 2016). This makes the direct task results appear weaker than those in the indirect task even when both tasks are probing the same information processing.

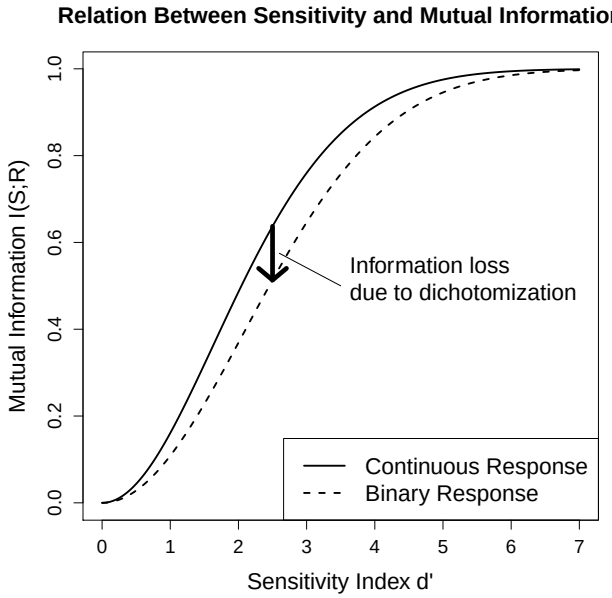

Figure 1.1: **Mutual Information is higher for continuous than binary responses in Signal Detection Theory (SDT).** In classical SDT, the signal is assumed to be binary ($S = \pm 1$) and the response is the signal corrupted by Gaussian noise ($R = S + N$, where $N \sim \mathcal{N}$). The sensitivity index d' indicates how far the two resulting Gaussian distributions are separated. The mutual information ($I(S;R)$, defined in Chapter 4) between signal S and these response R (solid line) is larger than the mutual information between signal and the dichotomized, binary response (dashed line): When reducing the continuous response to a binary prediction, information about the signal is lost.

This and other issues described in Chapter 2 lead to a fundamental problem with standard reasoning: It routinely infers a difference between conscious and unconscious tasks even in cases where there is no true difference. The problem with the standard reasoning is that the two tasks are not evaluated in a comparable manner. Instead, they are analyzed and evaluated separately with measures and methods that make the direct task appear weaker. The

conclusion that better processing was indicated by the indirect as compared to the direct task is therefore not established with the appropriate statistical methods. This is especially problematic when the conscious task is then also sampled with much fewer trials (as is often done; Vadillo, Konstantinidis, & Shanks, 2016). The apparent difference in favor of the unconscious task may simply be due to methodological issues. Thus, results from this standard reasoning are no reliable evidence for unconscious processing.

We demonstrate that a difference between the conscious and the unconscious task is erroneously inferred even when the amount of information processed in both tasks is exactly the same. This extends the work of Franz and von Luxburg (2015) who criticized the study of ten Brinke et al. (2014), where the standard reasoning was applied to conclude that unconscious lie-detection is superior to conscious lie-detection. The study by ten Brinke et al. also based their conclusion on the typical pattern of results: a weak performance when participants directly attempt to identify stimuli vs. a clear priming effect on reaction times. However, when Franz and von Luxburg (2015) reanalyzed the reaction time data of that study, they found no evidence for better unconscious identification of liars. Instead, participants seem to be able to consciously identify liars to the same extent that indirect measures reveal. Thus, there is no evidence for unconscious processing beyond consciously accessible processing. The same problem may have occurred in any study from the past decades that used the standard reasoning. With that, previous research has used too liberal methods when accepting evidence for unconscious processing. The standard reasoning they applied may have led to many such unwarranted conclusions about unconscious processing.

We scrutinize the methodology of these studies and develop an appropriate method. Our appropriate method equates evaluations in both tasks and compares directly how much information is processed in both. This way, the appropriate method avoids incorrectly concluding a difference between two incomparable measures. This prevents unwarranted claims about unconscious processing. We will also develop a variant of this appropriate method that can be used to reanalyze previous studies—without requiring access to the original data. With this, we investigate whether the conclusions about unconscious processing in the most influential studies from the past two decades were warranted or not.

Surprisingly, even though the reanalyzed studies claimed that, in their various settings, information was processed unconsciously, the appropriate method yields a different result. In many settings, there is no consistent evidence that human participants processed information beyond what they can consciously report. In general, we find no evidence against the idea that participants have

full conscious access to the processes evidenced by the indirect task results. Instead, evidence rather suggests that humans seem to have full introspective capabilities (Peters & Lau, 2015). This result has major implications on our theoretical understanding of the functional role of consciousness. Future priming studies seeking genuine evidence for unconscious processing must therefore abandon the standard reasoning and apply more appropriate methods, like the one developed here.

1.4 Statistical Aggregations in Group Decision Making

In the field of group decision making, studies take reports of the individuals and aim to predict what they will decide as a group. These studies use statistical aggregation methods to combine the individual reports into *simulated* group decisions (see examples in Hautz et al., 2015; Klein & Epley, 2015; Koriat, 2015; Kurvers et al., 2016; van Dijk et al., 2014). They do this before a real group discussion takes place. Then, the individuals convene as a group in an interactive group discussion and form a *real* group decision. The real group decision is then compared to the simulated decision. The goal of these studies is to find simulation methods that are either as consistent as possible to the real groups (Bahrami et al., 2010; Koriat, 2012a) or as good as possible (potentially outperforming real groups; Mannes, Soll, & Larrick, 2014; Litvinova, Herzog, Kall, Pleskac, & Hertwig, 2020). Such models can then be used to better understand or, perhaps, substitute real group discussions.

One popular example is Galton's ox (Galton, 1907; see also Wallis, 2014). During a fair, 787 visitors individually guessed the weight of an ox. The individual weight guesses deviated substantially (interquartile range of 74 kg). But when combining the individual guesses by taking the average yielded a good guess (only 4 kg off of the actual value). Averaging is a simple method to simulate what the group would have guessed had all 787 visitors been allowed to discuss. But an interaction in such a big group would have been impractical. Instead, replacing the real group interaction with simulation methods can save time and costs. The main question is then whether the simulation method, taking the average, adequately reflects what the real group would have decided. In many cases, simulation methods do not match what real groups do. Simulated decisions are typically worse than those produced by real groups (Bahrami et al., 2010; Birnbaum & Diecidue, 2015; Klein & Epley, 2015; Sniezek & Henry, 1989). In Galton's ox example, perhaps a real group—after a long deliberation period—would have derived a guess even closer to the

true value.

As in unconscious priming, Information Theory sheds light on the methods by guiding our attention to the information transmitted by the individual participants. In this case, again, mutual information measures not only how good participants' single weight guesses (point estimates) match the true value but also incorporates confidences. Human participants often not only know what decision they would make but also how confident they are in their decisions (Brenner et al., 1996; Fleming, Dolan, & Frith, 2012; Griffin & Tversky, 1992; Regenwetter et al., 2014). It seems natural that a real group discussion would incorporate these confidences in some way. For example, experts may receive a higher weight than laypeople (Bang et al., 2014). Thus, Information Theory prompts us to consider confidences when simulating group decisions.

Without invoking Information Theory, previous methods have acknowledged the role of confidences in group simulation methods. As an early example, Sniezek and Henry (1989) let participants estimate various quantities by asking participants to report subjective confidence intervals. Smaller confidence intervals correspond to more confident and larger intervals to less confident guesses (see also Sniezek & Henry, 1990 and Einhorn & Hogarth, 1978). Other studies have investigated participants' confidence ratings directly in terms of subjective probability ratings (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Griffin & Brenner, 2004). The typical result in these studies is that, for difficult tasks, participants are overconfident whereas, for easy tasks, participants are often underconfident. As a side note, this effect has been in disputed for methodological reasons (Fiedler & Krueger, 2012; Olsson, 2014) which are similar to those discussed regarding the infamous Dunning-Krueger effect (Kruger & Dunning, 1999, and see its methodological critique Campbell & Kenny, 1999; Nuhfer, Cogan, Fleisher, Gaze, & Wirth, 2016; Nuhfer, Fleisher, Cogan, Wirth, & Gaze, 2017). Nevertheless, the easy-hard effect survived even under scrutinized methodology because it is derived in a different way than the Dunning-Kruger effect. We will demonstrate one instance of this easy-hard effect Section 3.4. However, biases only degrade the additional information provided in confidence ratings but do not eradicate it. Thus, confidence ratings provide additional information and simulation methods almost always improve when taking them into account (Einhorn, Hogarth, & Klempner, 1977; Koriat, 2012a, 2012b; Litvinova et al., 2020; Mannes et al., 2014; Sniezek & Henry, 1989).

However, simulations often still fall short of real group decisions (Bahrami et al., 2010; Birnbaum & Diecidue, 2015; Klein & Epley, 2015; Sniezek & Henry, 1989). Real groups typically still outperform many simulation methods even when they take confidences into account. Thus, there is still a gap

between what real groups do and what current simulation methods suggest. It would seem as if the real groups have some elusive advantage over using statistical aggregations. For example, Klein and Epley (2015) found that real groups are better in identifying liars than statistically aggregating individual responses. To investigate this gap, we further look into how the information provided by participants' reports (guesses plus confidences) enters the simulated group decision. We find that most methods do not take the confidence ratings into account in the mathematically optimal way. For example, Koriat (2012a, 2012b) uses confidence ratings to select the most confident individual and simulates the group decision to be that of the most confident individual; Mannes et al. (2014) similarly selects a small group of the most confident individuals determining the simulated group decision from their reports.

We put forward the theoretically optimal method to incorporate individual's confidences into simulated group decisions. This method is known as Confidence Weighted Majority Voting (CWMV) in the mathematical literature (Grofman et al., 1983; Nitzan & Paroush, 1982). Even though many psychological studies on group decisions already incorporate confidence ratings, few do so with this background (we are only aware of Bahrami et al., 2010). We demonstrate how this method can be applied in a simple group decision experiment. There, we show that simulated group decisions based on this optimal method are as accurate as the real group decisions.

Even though CWMV produces the same accuracy as real groups in our experiment, real groups still deviated from simulations. To model real group processes, we adapt the method by introducing additional model parameters to capture two aspects of human group decision making. First, we show that real groups in our experiments exhibit an equality effect meaning that real groups tend to give more weight to unconfident group members and, conversely, less weight to confident group members compared to the theoretically optimal aggregation method. Second, real groups tend to be less confident than the aggregation method suggests. We thereby close the gap between real and simulated group processes and demonstrate the importance of considering confidences in the mathematically optimal way.

Our view on group decision making studies is similar to that of unconscious priming studies: In both cases, human participants give responses that later enter a comparison. In unconscious priming, individual responses from a direct task are compared to indirect measures. In group decision making, individual responses aggregated into simulated group decisions are compared to real group decisions. It is crucial that all the information human participants can express is captured and used appropriately. Otherwise, these comparisons are not proper and produce biased comparisons in favor of the indirect,

unconscious measures or the real group decisions. When the comparison is biased for methodological reasons, the true comparison of interest is occluded. Conclusions based on such biased comparisons are rendered invalid because a difference can simply be attributed to a failure of the method to adequately incorporate what human participants could have reported when allowed to. In both research topics, we discuss more appropriate methods to be used in future research to reduce methodological bias.

1.5 Ensemble Accuracy Bounds

While looking at group decision processes through the lens of Information Theory, we arrived at a third and more theoretical research question: How well does a group perform—given the performance of its members? In this third research topic, we prove bounds on the best- and worst-case accuracy of a group. This is relevant for group decision making in psychology and perhaps even more in machine learning (Bishop, 2006; Murphy, 2012). There, instead of human group members, algorithms produce individual classifiers that make predictions with certain accuracies. In many cases, a group of multiple classifiers is used to improve accuracy over that of any single individual classifier (Dietterich, 2000; Okun, Valentini, & Re, 2011; Schapire & Freund, 2013; Zhou, 2012). Such a group of classifiers is called an ensemble. It is not clear how much an ensemble's accuracy improves when multiple individual classifiers are combined. Except for simple cases, even when the accuracies of the individual classifiers are known, the ensemble accuracy is not.

On the contrary, it is known that even when two individual classifiers have exactly the same accuracy, they can contribute differently to the accuracy of an ensemble. Bartlett, Freund, Lee, and Schapire (1998) explained this via a concept called margins (see also Schapire & Freund, 2013, Chapter 5 and, for a discussion, Gao & Zhou, 2013; Koltchinskii, Panchenko, et al., 2002; L. Wang, Sugiyama, Yang, Zhou, & Feng, 2008). Margins represent how far away a decision is from a decision boundary, which is where the classifier would switch to a different decision. With large margins, classifiers make stable, reliable decisions whereas decisions with small margins are unreliable. Margins, therefore, correspond to confidences and convey measurable information.

We contribute by developing a model in which we formalize this explanation via *local* confidences—local in the sense that each prediction of a classifier comes with its own confidence instead of assigning a global confidence to all predictions of a classifier due to its overall quality. In this model, we compute the information of a classifier based on the distribution of confidences, which turns out to be a decisive factor for how much a classifier contributes to the

ensemble accuracy (Koltchinskii et al., 2002).

To illustrate this, consider the following example. Say, we want to know what the weather will be tomorrow, whether it will rain or not (cf. DeGroot & Fienberg, 1983). For simplicity assume that it rains in 50% of the days (or, equivalently, assume living in Glasgow, Scotland). We have two weather forecast channels that we can consult, A and B. Channel A and B are similar in that they both correctly predict tomorrow's weather in 80% of the cases. That is, in four out of five days, they predict correctly. But there is one difference. Channel A always just gives their prediction without further information. Since we know that 80% of the predictions are correct, there is always some uncertainty. Even if A says it will not rain, we cannot be sure and would always have to bring an umbrella. Channel B on the other hand adds meta-information in form of confidences about their daily predictions. On some days, B announces absolute certainty making this day's prediction 100% accurate. On other days, B admits a very low certainty such that we know not more about tomorrow's weather than if we had thrown a coin. These two cases are balanced such that B still produces 80% correct predictions in the long run: 60% of the predictions are made with absolute certainty and 40% are made with low certainty. This way, forecasts in three out of five days (60%) are certainly correct, and we know for sure that these will be correct. The remaining two days (40%) are randomly guessed and we know we cannot rely on them: Only half of these predictions, one out of two days, will be correct. In total, there will again be four out of five correct predictions from B because the certain three days are correct and half of the two random guessing days are correct. Thus, B also makes 80% correct predictions as A. But the advantage is that there will be some days on which channel B predicts no rain with absolute certainty and, on these days, we can leave our umbrella at home.

The above example shows two classifiers that have the same accuracy but convey different amounts of information. Channel A conveys less information than B because B adds meta-information by distinguishing how accurate each individual prediction is. This information is measurable. While A conveys 0.28 bit of information about tomorrow's weather, B conveys 0.60 bit. We will show in Chapter 4 that type A predictions convey the least possible information given their accuracy and type B predictions convey the most.

The surplus of information is relevant, especially in ensembles. When we want to consult multiple weather channels to increase accuracy, we can choose between multiple type A or multiple type B channels (a mixture is also conceivable, but not relevant for now). Say, we can either consult three channels of type A or, alternatively, three channels of type B. Assume these channels make independent predictions, that is, they make their own measurements and

forecasts instead of copying each others' result (in which case there would be nothing to gain from consulting multiple predictions). Based on simple probability calculations, consulting three type A channels will lead to an accuracy of approximately 90% in predicting tomorrow's weather. On the other hand, consulting three type B channels leads to an accuracy of about 97%. This is a striking difference of 7%-points in ensemble accuracy given that all channels individually have the same accuracy of 80%. Type B forecasts have an advantage in ensembles because they allow uncertain predictions to be dismissed. The predictive uncertainty is localized via the confidences by B. On the other hand, all predictions from A incorporate uncertainty making them less useful in ensembles.

We present proofs for tight lower and upper bounds for the ensemble accuracies. In the example above, this was the 90%-97% range of possible ensemble accuracies. Such bounds allow anticipating the ensemble accuracy of multiple individual classifiers. Moreover, they allow determining how many classifiers are necessary to guarantee a prespecified ensemble accuracy. Consider for example, that we require an accuracy of 95%. Assume we have a pool of individual classifiers that each predict independently with an accuracy of 80%. In principle, three of these classifiers are not enough to guarantee 95% ensemble accuracy because when they are of type A, they only achieve 90% as seen before. However, using our bounds we can derive that seven classifiers are guaranteed to be enough to achieve the desired ensemble accuracy.

We refine our results by taking into account the individual classifiers' information (the mutual information between their outputs and the target label that is to be predicted, e.g., whether it will rain or not). The previous bounds allow us statements like "Three independent classifiers with individual accuracy of 80% produce an ensemble accuracy between 90%–97%" and we can improve on that with statements like "Three independent classifiers with individual accuracy of 80% *and* information of 0.55 bit (close to the maximum of 0.60 bit for that accuracy) produce an ensemble accuracy between 94–97%." Thus, knowing not only the accuracy of the individual classifiers but also their information measure we can improve the ensemble accuracy bounds.

Furthermore, our bounds are constructive and yield guiding principles for selecting and building classifiers. Type B classifiers output predictions with very high confidences on some instances and very low confidences on others and we will therefore call them *specialists*. Type A classifiers always predict with an intermediate confidence making them *generalists*. Interestingly, even when classifiers are independent and do not coordinate their specializations, specialists have desirable properties. The independence assumption is important to understand the novelty of our results because it is obvious that (dependent)

specialists outperform generalists but it is quite surprising that independent specialists do so too, see Discussion on pp. 150–152.

Outlook

The main part of this thesis has three chapters, one for each research topic. In Chapter 2, we scrutinize a frequently used method—the standard reasoning—in unconscious priming studies. We explain the appropriate method and show that this standard reasoning produced misleading evidence. We then apply the appropriate method. In many situations, we find little to no evidence for unconscious processing contrary to previous claims. In Chapter 3, we put forward a method, CWMV, for simulated group decisions. In a simple setting, we use this theoretically optimal method to simulate real group decisions improving upon previous methods. We will also model how real groups deviate from this method. In Chapter 4, we develop a new model with local confidences that allows predicting the ensemble performance based on the performance of its members. These chapters are based on three manuscripts two of which (Chapters 2 and 3) are accepted by peer reviewed journals. Chapter 5 concludes with a discussion of our results and presents potential research directions building on our work. References and author contributions are provided in Chapter 6.

Chapter 2

Advancing Research on Unconscious Priming: The Indirect Task Advantage

Published as: Meyen, S., Zerweck, I. A., Amado, C., von Luxburg, U., & Franz, V. H. (2021). Advancing Research on Unconscious Priming: When can Scientists Claim an Indirect Task Advantage? *Journal of Experimental Psychology: General*. Advance Online Publication. https://doi.org/10.1037/xge0001065

> "The experimental work so far offers no convincing evidence that the human observer can respond differentially to stimulation that is of an intensity level below that at which a discriminated verbal report can be obtained. In other words, when adequate controls are provided, no response has been found to yield better absolute discrimination than a verbal response."
>
> — Eriksen (1960, p. 291)

Abstract

Current literature holds that many cognitive functions can be performed outside consciousness. Evidence for this view comes from unconscious

priming. In a typical experiment, visual stimuli are masked such that participants are close to chance performance when directly asked to which of two categories the stimuli belong. This close-to-zero sensitivity is seen as evidence that participants cannot consciously report the category of the masked stimuli. Nevertheless, the category of the masked stimuli can indirectly affect responses to other stimuli (e.g., reaction times or brain activity)—an effect called priming. The priming effect is seen as evidence for a higher sensitivity to the masked stimuli in the indirect responses as compared to the direct responses. Such an apparent difference in sensitivities is taken as evidence that processing occurred unconsciously. But we show that this "standard reasoning of unconscious priming" is flawed: Sensitivities are not properly compared, creating the wrong impression of a difference in sensitivities even if there is none. We describe the appropriate way to determine sensitivities, replicate the behavioral part of a landmark study, develop methods to estimate sensitivities from reported summary statistics of published studies, and use these methods to reanalyze 15 highly influential studies. Results show that the interpretations of many studies need to be changed and that a community effort is required to reassess the vast literature on unconscious priming. This process will allow scientists to learn more about the true boundary conditions of unconscious priming, thereby advancing the scientific understanding of consciousness.

Research on consciousness and its cerebral substrates has far-reaching implications and received substantial attention in recent years (Michel et al., 2019). A driving factor comes from reports that masked stimuli that are not consciously perceived can nevertheless affect behavioral responses and brain activity (Kouider & Dehaene, 2007; van den Bussche, van den Noortgate, & Reynvoet, 2009). The exciting claim here is that unconscious processing might be more than a mere residue of conscious processing and may be performed by different neuronal processes than conscious processing. Such results impact current theories about the functional role of consciousness (Dehaene, Lau, & Kouider, 2017; Kouider & Dehaene, 2007; van den Bussche et al., 2009; Sklar et al., 2012; Hassin, 2013), might suggest parallel neuronal routes for unconscious vs. conscious processing (Morris, Öhman, & Dolan, 1999), and might support theories of superior unconscious processing (Custers & Aarts, 2010; Dijksterhuis et al., 2006; ten Brinke et al., 2016).

Here, we scrutinize one of the most frequently used approaches in this field. We show that the **standard reasoning** in the dissociation paradigm (Hannula et al., 2005; Holender, 1986; Schmidt & Vorberg, 2006; Simons et al., 2007) is flawed for mathematical reasons. It fails to provide meaningful interpretation

of the data, and needs to be replaced by an *appropriate analysis*. Because many studies have used the standard reasoning, a large body of literature needs reassessment. This has the potential to drastically change our views on unconscious processing and its neuronal underpinnings. The fallacy we expose affects a wide range of research areas because the standard reasoning has been employed on such diverse topics as, for example, unconscious processing of semantic meaning (Dehaene et al., 1998), motivation (Pessiglione et al., 2007), emotion (Morris, Öhman, & Dolan, 1998), cognitive control (van Gaal et al., 2010), and detection of lies (ten Brinke et al., 2014).

To assess how seriously the literature is affected, we proceeded in three strands: (a) We replicated the behavioral part of a landmark study (Dehaene et al., 1998) and showed that the appropriate analysis of the data does not support unconscious priming (in contrast to the claims of the original study). (b) We developed statistical methods to reanalyze published studies based on the reported t and F statistics (because access to the full trial-by-trial data is often lacking). We validated this approach by showing that our reanalysis of the published data of Dehaene et al. (1998) is consistent with the results of our replication. (c) We used our methods to reanalyze 15 highly influential studies (with a total of 3277 citations in Web of Science). Even though all these studies used the standard reasoning to infer unconscious processing, their data tell a different story.

2.1 The Standard Reasoning of Unconscious Priming Implies an ITA

As a typical example for a study using the standard reasoning, consider the study by ten Brinke et al. (2014) who reported that humans can detect liars better unconsciously than consciously: "[T]he unconscious mind identifies and processes cues to deception ... more efficiently and effectively than the conscious mind." (p. 1104). In the following, we will describe the specifics of this study as well as the general aspects that are typical for studies using the standard reasoning.

Participants of ten Brinke et al. (2014) first watched videos of suspects who were either lying or telling the truth. Then participants performed two tasks: The direct and the indirect task. These tasks were supposed to measure conscious and unconscious lie detection, respectively.

In the **direct task**, participants judged which suspects had been lying or telling the truth. Participants performed poorly with an accuracy of only 49.62%-correct (with chance level being 50%), which was taken by ten Brinke

et al. (2014) as evidence that participants could not consciously detect liars with more than a poor sensitivity. In the same way, studies using the standard reasoning typically let participants directly discriminate stimuli belonging to one of two categories (Figure 2.1). Participants' performance—measured by the proportion of correct responses or by the sensitivity index, d', from Signal Detection Theory (Green & Swets, 1988)—is typically found to be close to chance level. This result is then taken as evidence that conscious discrimination of the presented stimuli is poor at best.

In the **indirect task** of ten Brinke et al. (2014), participants categorized target–words, such as "deceitful" or "honest", into two categories: lying or truth–telling. Before each target–word, a masked picture of one of the suspects was briefly presented in order to affect (or "prime") the responses to the target words (therefore those masked stimuli are often called the "primes"). ten Brinke et al. found that participants' reaction times (RTs) to the target words were faster when the primes were congruent with the targets (e.g., the picture of a lying suspect before a lie–related word) than when the primes were incongruent with the targets. That is, ten Brinke et al. (2014) found a congruency effect between primes and targets in the indirect task. In the same way, studies using the standard reasoning typically employ an indirect task attempting to find such congruency effects (Figure 2.1). These congruency effects could be on RTs (as in the case ten Brinke et al., 2014), but also on other behavioral responses (e.g., skin conductance) or neurophysiological measures (e.g., in EEG or fMRI).

Taken together, ten Brinke et al. (2014) found the typical pattern of results for the unconscious priming paradigm: (a) a poor accuracy, or sensitivity in the direct task and (b) a clear congruency effect in the indirect task. Based on this pattern, they concluded that participants' indirect task revealed more accurate lie detection than the direct task: "[I]ndirect measures of deception detection are significantly more accurate than direct measures" (p. 1098, Abstract). In the same way, studies using the standard reasoning infer from such a pattern of results better sensitivity for the primes in the indirect task than in the direct task (Figure 2.2). We dubbed this situation the **indirect task advantage**, or short **ITA**. It is important to note that the claim of an ITA is, in this phase of the reasoning, independent of any considerations about conscious or unconscious processing. We call this descriptive phase of the standard reasoning **Step 1**.

In **Step 2** of the standard reasoning, ten Brinke et al. (2014) used the presumed ITA to conclude superior unconscious processing: "[A]lthough humans cannot consciously discriminate liars from truth tellers, they do have a sense, on some less-conscious level, of when someone is lying" (p. 1103). The

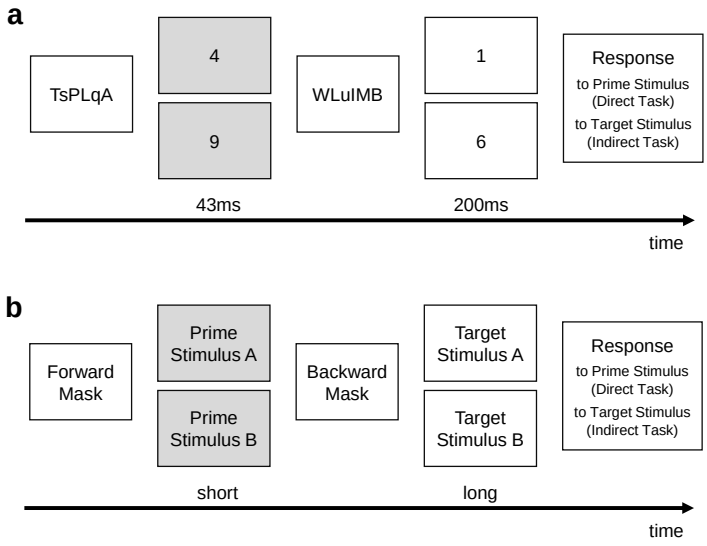

Figure 2.1: **Typical study design to infer an indirect task advantage (ITA). (a) Example study:** The study of Dehaene et al. (1998) is a prototypical example for unconscious priming with number stimuli. In each trial, a masked *prime* stimulus is presented for a short duration followed by a well visible target stimulus. In the direct task, participants discriminated the primes and performance was close to chance level. In the indirect task, participants discriminated the *target* stimuli by deciding whether they were smaller or larger than the number 5. Reaction times (RTs) were faster and lateralization of brain activity was larger when prime and target stimuli were congruent (both smaller or both larger than 5) than when they were incongruent (one larger one smaller). Dehaene et al. (1998) followed the standard reasoning to infer a higher sensitivity for the primes in the indirect task than in the direct task (i.e., an ITA) and conclude that the primes were processed in the absence of conscious awareness. **(b) General design:** In general, prime and target stimuli each come from one of two categories, A or B. In the direct task, participants discriminate the prime (e.g., guess whether it is from category A or B) with a poor sensitivity. In the indirect task, the same stimuli are presented and participants now discriminate the target. In this task, the prime is shown to influence responses resulting in faster RTs for congruent (A–A, or B–B) than incongruent trials (incongruent: A–B, or B–A). From this pattern of results, the standard reasoning infers an ITA (cf. Figure 2.2).

authors thereby followed the typical assumption that direct and indirect tasks measure conscious and unconscious processing, respectively. Based on the supposed ITA from Step 1, these assumptions lead to the typical conclusion that participants processed the category of the masked stimuli better unconsciously than they can consciously report.

The standard reasoning is summarized for example by Dell'Acqua and Grainger (1999): "The present work follows the tradition of providing evidence for a dissociation between direct and indirect effects of unconsciously

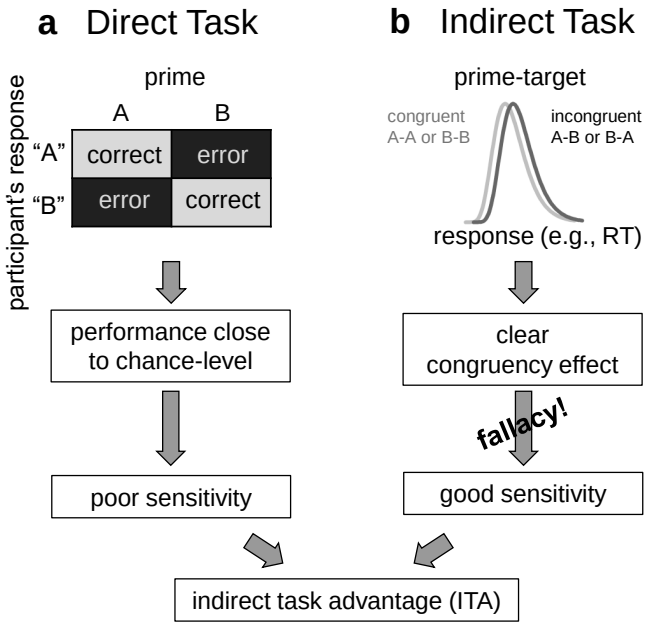

Figure 2.2: **Standard reasoning to infer an indirect task advantage (ITA). (a)** In the *direct task*, the standard reasoning infers from close-to-chance performance that there was poor sensitivity for the primes, if any at all. **(b)** In the *indirect task*, the standard reasoning infers from a clear congruency effect that sensitivity was relatively good. Based on this pattern of results the standard reasoning makes two inference steps: In Step 1, it incorrectly infers that participants' responses in the indirect task were more sensitive to the primes than responses in the direct task (ITA). In Step 2, it attributes this difference to unconscious processing. However, already Step 1 of this reasoning is flawed because a clear congruency effect does not necessarily indicate good sensitivity. It could be caused by a sensitivity as poor as (or even worse than) the sensitivity in the direct task! Because Step 1 is independent of any (sometimes contentious) assumptions about conscious vs. unconscious processing, our critique is also independent of any such assumptions.

presented stimuli (Greenwald, Klinger & Schuh, 1995; Draine & Greenwald, 1998). More specifically, null effects are sought in direct measures (i.e. where subjects respond directly to the unconsciously presented stimuli) accompanied by non-null indirect effects (i.e. priming effects)" (p. B2). For further description of the standard reasoning see also Merikle (1992) and Simons et al. (2007). Even though some studies may not state an ITA as explicitly as shown here, it is nevertheless necessarily implied when claims about unconscious processing are made because Step 1 is a necessary condition for Step 2.

But note that the standard reasoning infers better sensitivity in the indirect task than in the direct task (i.e., an ITA) without ever calculating sensitiv-

ity (or accuracy) in the indirect task to compare against that in the direct task. For example, ten Brinke et al. (2014) only demonstrated a congruency effect on RTs. **However, if this congruency effect indicated accurate unconscious lie detection, we should be able to use the RT data to determine which of the suspects were lying with a higher accuracy than in the direct task.** Otherwise the congruency effect does not truly provide evidence for better accuracy in the indirect than in the direct task (i.e., for an ITA).

Because ten Brinke et al. (2014) laudably followed an open-data policy, Franz and von Luxburg (2015) were able to reanalyze how much evidence the RT data truly provided for better accuracy in the indirect than in the direct task. To assess this, they determined statistically optimal classifiers, used the RT of each trial in the indirect task to classify ("predict" in the nomenclature of statistical learning) which of the suspects were lying, and found the accuracy in the indirect task to be only at 50.6%-correct ($SEM = 0.3\%$; see below for more details on the methods used). This value is very similar to—and not significantly different from—the accuracy in the direct task (which was 49.62%-correct; $SEM = 1.4\%$). Therefore, ten Brinke et al.'s inference in Step 1 was flawed: Their data did not provide evidence for better accuracy in the indirect than in the direct task. In our words, there was no evidence for an ITA. Because the existence of an ITA in Step 1 is a necessary condition for Step 2 of the standard reasoning, inferences about unconscious processing were not warranted.

In the following section, we show in detail why claiming an ITA based on the standard reasoning is flawed. Note, that our critique focuses on how an ITA is established in Step 1 and is therefore independent of any assumptions about conscious vs. unconscious processing, which are relevant only in Step 2 and for which different, sometimes contentious approaches exist (e.g., Eriksen, 1960; Erdelyi, 1986; Holender, 1986; Reingold & Merikle, 1988, 1990; Schmidt & Vorberg, 2006). We avoid these discussions by focusing on an empirical investigation of Step 1 which makes our critique very general.

2.2 The Standard Reasoning is Flawed and Fails to Provide Evidence for an ITA

The standard reasoning is intuitively very appealing, which seems to be one reason for its popularity. The colloquial version of the arguments to infer an ITA in Step 1 goes like this: "Participants have a very hard time to discriminate the masked stimuli in the direct task. They are very close to zero sensitivity

and usually not significantly above chance. Nevertheless we find clear and highly significant congruency effects in the indirect task. Therefore, it seems obvious, that the indirect task responses are more sensitive to the masked stimuli than the direct task responses."

However, this intuition is misguided. To see this, consider what happens if we increased the number of observations (number of participants or trials). The poor sensitivity in the direct task (Figure 2.2a) will only be measured more precisely but will still be poor. In contrast, the congruency effect in the indirect task (Figure 2.2b) becomes clearer because it is based on the difference between congruent and incongruent condition means: With more observations, the variability of the two means becomes smaller, such that the difference between them becomes clearer. Therefore, a clear congruency effect can be generated by a good underlying sensitivity (corresponding to, say, $d' = 5$ or 99%-correct) but it can also be generated by a very poor sensitivity (say, $d' = 0.05$ or 51%-correct). In cases where the sensitivity of the indirect task is as poor as in the direct task, there is no ITA and further interpretations about unconscious processing are unwarranted. Not recognizing this is the **main fallacy of the standard reasoning**. We demonstrate this problem by using a toy example.

Toy Example With Baby Weights

To illustrate the problem of the standard reasoning, consider an example in which responses in the direct and indirect tasks are based on the *exact same* underlying sensitivity. Nevertheless, the standard reasoning would erroneously infer that responses in the indirect task were *more* sensitive than responses in the direct task (i.e., an ITA).

Consider participants measured the birth weights of newborn girls (category A) and boys (category B), such that they only knew the weight of the babies but not the biological sex. This would be all the information participants had in both, direct and indirect, tasks.

In the direct task, participants would use this weight information to guess whether a baby is a girl or a boy (newborn girls weigh a little less than boys). Due to the large overlap between the weight distributions (Figure 2.3a), participants would be correct in only approximately 55% of the cases even when using an ideal decision criterion. This corresponds to a poor performance that is close-to-chance level (50%). Following the standard reasoning, an experimenter would correctly infer a poor sensitivity in this direct task (Figure 2.2a).

In the indirect task, participants would simply report the numerically measured weight of the babies. The experimenter would *average* those responses across groups of baby girls and boys and would calculate the difference of the

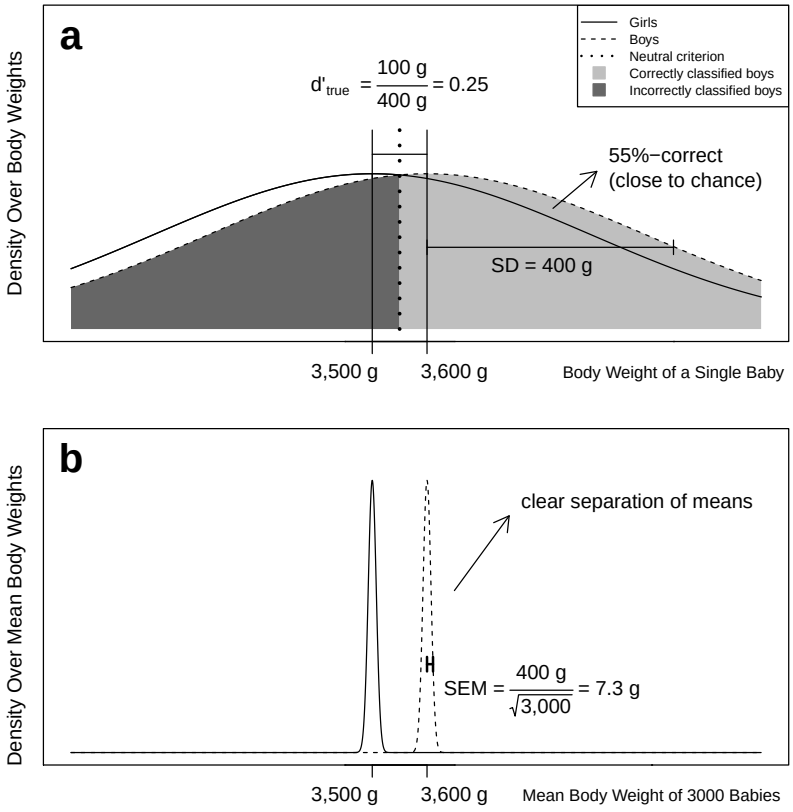

Figure 2.3: **Toy-example demonstrating fallacy of standard reasoning.** We show that even when responses in the direct and indirect tasks are based on the *exact same* information the standard reasoning would nevertheless infer an indirect task advantage (ITA): a *higher* sensitivity in the indirect as compared to the direct task. Consider participants of a hypothetical experiment measured the birth weight of babies but did not know the babies' sex. **(a)** In the *direct task*, participants used the weight of an individual baby to guess whether it is a girl or a boy. The weight distributions overlap heavily such that sensitivity would be poor ($d'_{\text{true}} = 0.25$; corresponding to 55%-correct). **(b)** In the *indirect task*, participants responded by simply stating the measured weights. The experimenter would average those responses across many trials (e.g., across 3000 girls and 3000 boys). The resulting group means are much less variable than the individual weights such that the experimenter would obtain a clear difference between the two group means (this corresponds to a clear congruency effect in the priming paradigm). Based on this result, the standard reasoning would erroneously infer that participants had relatively good sensitivity about whether a baby was a girl or a boy in the indirect task—better than in the direct task. That is, the standard reasoning would infer an ITA even though the *exact same* information created the responses in both tasks. Weight–data based on Janssen et al. (2007).

mean responses to those two groups. With increasing group sizes, the experimenter would eventually find a clear difference (corresponding to the clear

congruency effect in the priming paradigm). Figure 2.3b shows this for 3000 observations in each group, which is a typical number of observations in the indirect task (e.g., when 10 participants perform 300 trials in each condition, the number of observations per condition is $10 \times 300 = 3000$). Following the standard reasoning, the experimenter would incorrectly infer a good sensitivity in this indirect task (Figure 2.2b).

Here is the catch: The standard reasoning would incorrectly interpret this pattern of results as evidence for better sensitivity in the indirect task than in the direct task (i.e., for an ITA). However, this inference is wrong because participants gave responses in both tasks based on exactly the same information: In both tasks they knew only the weight of the babies. The illusion of an ITA is generated by the different data-analysis strategies of the experimenter in the two tasks and by the fact that the experimenter never attempted to estimate the sensitivity in the indirect task.

Further Details on the Standard Reasoning

We have shown that the standard reasoning is flawed because it infers an ITA in Step 1 even when there is none. The problem is that the standard reasoning calculates two very different things in the direct and indirect tasks: In the direct task, it calculates how well each stimulus can be classified on a trial-by-trial level. In the indirect task, it assesses whether there is a difference in mean responses. These are two very different things and it is a priori to be expected that the sensitivity in single trials can be poor while mean responses can nevertheless be clearly separated between the two categories given enough trials. A more appropriate analysis to determine whether there is an ITA would need to estimate sensitivities in both tasks and compare them. Before we present such an analysis, we want to first discuss some details of the standard reasoning.

True Zero-Sensitivity in the Direct Task

Consider that the true sensitivity in the direct task were known to be exactly zero and that there were at the same time a clear congruency effect in the indirect task. This ideal situation is typically sought—but typically not fully achieved—in the dissociation paradigm (Schmidt & Vorberg, 2006; Hannula et al., 2005; Simons et al., 2007). In this case (and only in this case), the standard reasoning would be justified in claiming that responses in the indirect task were somehow more sensitive than responses in the direct task. This is so, because a positive (larger than zero) sensitivity—even if it is minute—is required to produce a congruency effect and therefore the indirect task sensitivity must be

larger than zero. However, there are a number of problems with this scenario: (a) It is unrealistic. Typically, studies either find some small, residual sensitivity in the direct task or they do not find a congruency effect (Zerweck et al., 2021). (b) One cannot be certain of a true zero sensitivity. Instead, sensitivity in the direct task always needs to be measured (and is therefore affected by measurement error). Thus, we would still need to establish that the sensitivity in the indirect task is indeed larger than that in the direct task (e.g., by a significance test on the difference). (c) The sensitivity in the indirect task could still be so low, that it would be close–enough to the zero sensitivity of the direct task to not allow for strong conclusions (e.g., consider a sensitivity that corresponded to 50%-correct in the direct task and to 51%-correct in the indirect task).

Significance Testing vs. Bayesian methods

Until now, we purposefully did not talk about statistical significance testing because we wanted to focus on the main fallacy of the standard reasoning. Because significance testing and its applications have been heavily—and often rightfully—criticized since the very inception of the concept (Boring, 1919; Morrison & Henkel, 1970; Dienes, 2011; Cumming, 2014), it might be tempting to attribute the main fallacy of the standard reasoning also to significance testing. However, the problem of the standard reasoning is not so much that the statistical tools were wrong, but that the wrong statistical question is asked for the indirect task: The standard reasoning asks whether there is a true difference in means between congruent and incongruent conditions. However, the correct question to ask would be what the sensitivity in the indirect task is and whether this sensitivity is higher than in the direct task (such that an ITA can be concluded). Therefore, it would not help to simply replace the frequentist significance testing by Bayesian methods. Because researchers are interested in establishing an ITA (i.e., a difference in sensitivities) it does not suffice to evaluate both tasks in isolation. We must test directly for a difference in sensitivities between the two tasks. Failure to do so can lead to serious errors no matter whether we used significance testing (cf. Appendix B of Franz & Gegenfurtner, 2008, and Nieuwenhuis, Forstmann, & Wagenmakers, 2011) or Bayesian methods (cf. Supplement 2.6.6, and Palfi & Dienes, 2020).

Direct Task is Typically Underpowered

An additional problem in the application of the standard reasoning arises from the widespread use of seriously underpowered direct tasks (Buchner & Wippich, 2000; Vadillo et al., 2016; Vadillo, Linssen, Orgaz, Parsons, & Shanks,

2020). When the direct task is sampled with fewer participants and trials than the indirect task (as is often the case), a non-significant direct task result may not indicate that the true sensitivity is close to or exactly zero but rather that statistical power is low. Moreover, participants are required to give binary responses in the direct task in contrast to the continuous measures in the indirect task (e.g., RTs). Since participants have some continuous sense (confidence) about their responses (Rausch, Hellmann, & Zehetleitner, 2018; Zehetleitner & Rausch, 2013), the binary response format forces them to discard this information (Cohen, 1983), which further decreases the statistical power in the direct task. Therefore, even if the same sensitivity underlies responses in both tasks, it is a priori to be expected that the direct task produces less often significant results than the indirect task.

2.3 Appropriate Analysis: Calculate Sensitivities and Test for a Difference

We have shown that the standard reasoning is flawed and that researchers must compare sensitivities of both tasks if they want to infer an ITA. In this section, we describe more appropriate analyses. First, we assume that trial–by–trial data are available (this analysis was used by Franz & von Luxburg, 2015). Then we describe our newly developed method to reanalyze studies when only summary statistics are available. For detailed mathematical derivations see the online supplementary materials.

In deriving our methods, we unavoidably were confronted with degrees of freedom when choosing the details of our analysis strategy. In these cases, we chose strategies that favored finding an ITA. That is, we followed a **benefit-of-the-doubt** approach, thereby increasing the chances of confirming an ITA. We adopted this approach because we are criticizing a large body of literature. Therefore, it seemed necessary and reasonable to adopt such a liberal bias in confirming ITAs (and thereby being conservative in our critique) at this stage of the scientific discussion. It is understandable that researchers who have spent years using the standard reasoning might be reluctant to accept our arguments if our methods were too restrictive. This approach makes our arguments even stronger when we nevertheless do not find evidence for ITAs.

Sensitivity Comparison When Trial-By-Trial Data are Available

The appropriate method directly compares sensitivities in the direct and indirect tasks. Different than the standard reasoning, the appropriate analysis equates analysis steps for both tasks such that the calculated statistics are comparable. Then, a test of the difference between the two tasks is applied. Similar approaches have been used in previous—albeit very few—studies (Dulaney & Eriksen, 1959; Klotz & Neumann, 1999; Kunst-Wilson & Zajonc, 1980; Schmidt, 2002; Franz & von Luxburg, 2015) in accordance with the long standing (but often ignored) request for both tasks to be measured using the "same metric" (Reingold & Merikle, 1988).

In both tasks, we compute d' using Signal Detection Theory (Green & Swets, 1988) and then test for a difference between them. In the **direct task**, participants typically classify the primes in each trial and a d' value is often already reported by the studies using the standard reasoning. In the **indirect task**, however, the standard reasoning computes a congruency effect on continuous measures (e.g., RTs or brain activity as measured by EEG or fMRI). For a proper comparison, we have to transform these continuous measures into classifications (predictions) for each trial. There are different ways to achieve this. We suggest to use the optimal classifier for the given setup. This gives the indirect task the best possible performance and increases the chances of finding an ITA following the **benefit-of-the-doubt** approach.

Which classifier is best? We have shown that under typical conditions with equal number of congruent and incongruent trials, the median-split classifier is optimal (see Supplement of Franz & von Luxburg, 2014 for details and proof). The classification proceeds as follows: For each participant, we determine the median RT and classify ("predict" in the nomenclature of statistical learning) all trials with smaller RTs as congruent, and trials with larger RTs as incongruent. Then, we compare these classifications to the true labels (congruent/incongruent) evaluating for each trial whether the classification was correct or not, and we then compute a d' value as in the direct task. Finally, we compare the d' values of the direct and indirect task and test for a difference.

Some details: (a) Instead of computing d' values, the analysis could also be based on %-correct values. Assuming a neutral observer predicting both categories equally often in the direct task, both approaches produce the same results and we later report both measures to foster intuition. (b) Dichotomization of the continuous, indirect measures will result in a loss of information (Cohen, 1983). However, the direct task also requires participants to give binary responses. Converting indirect task responses into a binary response

format using our median split approach only equates this dichotomization to make responses in both tasks comparable. (c) We classify the trials of the indirect task according to the labels congruent/incongruent and not according to the prime category A/B, as is typically asked in the direct task. This is so because studies typically find a congruency effect between prime and target (and not a mere effect of the prime being in category A or B). For a comparison to the direct task, we would ideally transform the congruency classification into a classification of the prime category (A vs. B). For simplicity, we assume an optimal transformation here (without errors). This is plausible, because the target stimuli are typically fully visible to the participants, such that errors are rare. Again, our approach increases the chances of finding an ITA following the benefit-of-the-doubt approach.

Sensitivity Comparison When Only Summary Statistics are Available

Because the standard reasoning to infer an ITA is flawed, many already published studies on unconscious priming need reassessment. However, the appropriate analysis as described in the previous section would require full trial-by-trial data. Unfortunately, trial-by-trial data can be difficult or impossible to obtain for published studies (Wicherts, Borsboom, Kats, & Molenaar, 2006). For the older—but nevertheless influential—studies, those data might not even exist anymore. Therefore, we developed an approach that allows to estimate the results of the appropriate analysis without access to trial-by-trial data and solely based on the typically reported statistics from the standard reasoning. Here, we sketch the central approach of this analysis; details are given in Supplement 2.6.3.

The overall aim of this reanalysis is, again, to estimate sensitivities for the direct and indirect tasks (i.e., to either calculate d' from Signal Detection Theory or %-correct assuming a neutral observer). The direct task typically already provides d' or %-correct values. In the indirect task, studies typically report t or F values from a repeated measures design for the congruency effect. In this design, we show how F values can be translated to t values. We then derive an estimator for the underlying sensitivity that takes the form of a constant c_{N,K,q^2} multiplied onto the reported t value. This constant will include the number of participants N and trials K from the indirect task because t values become larger the more observations are made. Additionally, because this reanalysis can only use the reported statistics, one free parameter needs to be estimated: the ratio of between- vs. within-subject variances, which we denote by q^2. We estimated this parameter based on (a) our own

replication experiment (b) a literature review, and (c) extensive simulations (see Supplement 2.6.4). By assuming the largest plausible value for q^2, we again maximize the estimated sensitivity, d', in the indirect task and therefore increase the likelihood of confirming an ITA. Here, we again follow the benefit-of-the-doubt approach.

2.4 Replication of a Landmark Study Finds no ITA

We are now equipped with the appropriate tools that allow us to analyze typical settings and tasks that have been investigated in the context of unconscious priming. In this section, we will focus on one highly influential study on unconscious semantic priming of numbers (Dehaene et al., 1998). We will first describe the study and how its conclusions depend crucially on the flawed standard reasoning. Then, we will describe a replication experiment of the behavioral part of this study and analyze the trial-by-trial data. In the next section, we will then reanalyze the published results of this and other studies (15 in total). Overall, we will conclude that the results of our replication are similar to those of the original study. Both, our replication and our reanalysis of the original study, give reason to seriously doubt the existence of an ITA, questioning the authors' interpretation in the original study.

Dehaene et al. (1998) were interested in the question of whether the semantic meaning of numbers can be processed outside conscious awareness. They employed a prototypical priming experiment with stimuli shown in Figure 2.1a and applied the standard reasoning: In the direct task, participants discriminated features of masked numbers and performed poorly ($d' = 0.2$; corresponding to 54%-correct). In the indirect task, participants were again presented with the masked numbers (now serving as primes), but decided whether subsequent target numbers were smaller or larger than five. Participants responded approximately 24 ms faster when prime and target were congruent (both larger or smaller than five) than when they were incongruent (one smaller and one larger than five). Similar congruency effects were found for brain activity in EEG and fMRI (i.e., larger lateralization of brain activity in congruent than incongruent trials).

Dehaene et al. (1998) interpreted these results according to the standard reasoning: In **Step 1**, they inferred an ITA. That is, higher sensitivity in the indirect task than in the direct task: "[participants] could neither reliably report [the prime's] presence or absence nor discriminate it from a nonsense string [...] Nevertheless, we show here that the prime is processed to a high

cognitive level." (p. 597). In **Step 2**, they argued that "the prime was unconsciously processed" (p. 597) because participants were at chance performance in the direct task. Overall, they concluded: "By showing that a large amount of cerebral processing, including perception, semantic categorization and task execution, can be performed in the absence of consciousness, our results narrow down the search for its cerebral substrates" (p. 599). In short, Dehaene et al. (1998) employed a prototypical version of the standard reasoning to infer an ITA and unconscious processing exactly as described above. To assess the validity of these claims, we first replicate the behavioral part of that study, later we will reanalyze the published data.

Disclosures

Data, Materials, and Online Resources

The experimental material, data and the scripts for the analyses reported in this article have been made available on the Open Science Framework (OSF), at `https://osf.io/kp59h` (Meyen, Zerweck, Amado, von Luxburg, & Franz, 2020). We also provide an online tool to apply our reanalysis to other data at `http://www.ecogsci.cs.uni-tuebingen.de/ITAcalculator/`.

Reporting

We report how we determined our sample size, all data exclusions, and all measures in the study.

Methods

Twenty-four volunteers participated in our study (13 female, 5 left-handed, age range: 19–27 years; $M = 21.5$, $SD = 1.9$). All had normal or corrected-to-normal vision, signed written informed consent and were naive to the purpose of the experiment. In the original study by Dehaene et al. (1998), six and seven participants took part in the first and second direct task, respectively, and 12 participants took part in the indirect task.

We took great care to make stimuli and timings as similar as possible to those of the original study. Each trial consisted of: fixation cross (417 ms), forward mask (67 ms), prime (42 ms), backward mask (67 ms), and target (200 ms). In the original study, those values were: forward mask (71 ms), prime (43 ms), backward mask (71 ms), and target (200 ms; cf. Figure 2.1a). Slight differences in timing are due to slightly different refresh rates of the

monitors used. The prime duration of 43 ms was chosen by the original authors because it was the longest duration that produced non-significant results in the direct tasks. Primes and targets were numbers (1, 4, 6 or 9) that were either presented as digit (e.g., "1") or word (e.g., "EINS"; German for "ONE"). The original study used the same numbers in English, a follow–up used French (Kouider & Dehaene, 2009). As in the original study, primes and targets could be congruent (both smaller or both larger than 5) or incongruent (one smaller, one larger). Masks were composed of seven randomly drawn characters from {a-z, A-Z} mirroring the original study's masks. Participants were seated in front of a monitor (VIEWPixx /3D, VPixx Technologies Inc., Montreal, Canada), effective refresh rate 120 Hz at a viewing distance of approximately 60 cm in a sound- and light-protected cabin. In the original study, the monitor was a cathode-ray tube (CRT) with a refresh rate of 70 Hz. Stimuli were presented centrally as white text (69 cd/m^2; character height: 1°; width: 0.5° visual angle; font: Helvetica) on a black background (0.1 cd/m^2). These luminance values were not specified in the original study so that we chose the most plausible settings for our experiment.

In the direct task, participants classified whether the prime was smaller or larger than five. We used this particular task because the original authors argued in a subsequent study that it is "better matched with the [indirect] task" (Naccache & Dehaene, 2001b, p. 227). In the original study by Dehaene et al. (1998), two direct tasks were used, that produced similar results: In their first direct task, the prime stimulus was omitted in some trials and participants had to discriminate their presence vs. absence. In the second direct task, the prime stimuli were replaced by random letter strings and participants had to discriminate between numbers vs. random strings.

In the indirect task, participants decided as quickly as possible whether the target was smaller or larger than five; as was the case in the original study. Each participant performed 256 trials per task, preceded by 16 practice trials in each task. In contrast, in the original study, participants performed only 96 and 112 trials in the first and second direct tasks, respectively, and 512 trials in the indirect task.

We repeated indirect task trials with RTs that were too slow (> 1 s) or too fast (< 100 ms). The original study also rejected too slow trials (> 1 s) but was more restrictive in terms of fast trials: They rejected with RT < 250 ms. However, we only found 8 out of 6144 trials in our data to be above 100 ms but below 250 ms so that we obtained very similar results when applying their criterion. The indirect task was performed before the direct task (as is common practice in this paradigm) to prevent participants attending to the prime in the indirect task. In the original study, the direct and indirect tasks were

performed by different groups.

The number of participants and trials were chosen to produce a statistical power of above 95% to find a difference between sensitivities and confirm an ITA if it is there (see Supplement 2.6.2). For this power estimation, we assumed a true sensitivity of $d'_{\text{true, direct}} = 0$ in the direct task vs. $d'_{\text{true, indirect}} = 0.25$ in the indirect task (values based on our reanalysis of Dehaene et al., 1998, see below). To our knowledge, the original study did not perform a power analysis. A post hoc power analysis revealed that the original study had a statistical power of only 46% to find an ITA using the appropriate analysis (again, assuming $d'_{\text{true, direct}} = 0$ and $d'_{\text{true, indirect}} = 0.25$). This low power is due to a small number of direct task participants and trials.

Results and Discussion

Our analysis proceeded in two strands: First, we perform the traditional analysis which forms the basis for the standard reasoning. Second, we perform the appropriate analysis.

Standard Reasoning

The direct task sensitivity was $d' = 0.26$ ($SD = 0.27$), $t(23) = 4.68$, $p < .001$, corresponding to an accuracy in prime identification of $M = 54.87\%$-correct ($SD = 4.9$, $t(23) = 4.82$, $p < .001$). This is exactly in the range of sensitivities reported in the original study's direct tasks ($d' = 0.3$ in the first and $d' = 0.2$ in the second direct task). For a graphical depiction of these results, compare the bars corresponding to the direct tasks in Figures 2.4a and 2.4b.

Note that, in contrast to the original study, our direct task sensitivity is significantly above zero. This is so, because we sampled much more participants and trials than in the original study. Therefore, we had much higher statistical power. To simulate the lower power of the original study, we discarded data from participants and trials to match the same number of observations as in the original study: We kept only the first $N = 7$ participants and the first 112 trials of each participant. This leads to a non-significant result, $d' = 0.31$ ($SD = 0.39$), $t(6) = 2.06$, $p = 0.085$, as was the case in the original study. Therefore, it is plausible that it was the low statistical power in the original study (and not the sensitivity being exactly at zero) that was the reason for the nonsignificant result in the direct task of the original study.

In the indirect task, the congruent condition yielded faster RTs ($M = 445$ ms, $SD = 42$) than the incongruent condition ($M = 457$ ms, $SD = 37$), resulting in a clear and highly significant congruency effect of $M = 12$ ms

($SD = 11.8$), $t(23) = 4.95$, $p < .001$. That is, we found a highly significant congruency effect on RTs, as did the original study.

There is one potential caveat here: The congruency effect in the original study was larger than that in our replication (24 ms vs. 12 ms, respectively). However, we will show that sensitivities in our replication and our reanalysis of the original study are very consistent, see Figures 2.4a and 2.4b. This can be explained by larger trial-by-trial variability in the original study counteracting the larger RT effect: The original study, despite using 512 trials per participant, observed $SD = 13.5$ while we observed $SD = 11.8$ in our replication with only 256 trials per participant. Generally, more trials per participant should make individual RT effects more precisely measured. Thus, the standard deviation in the original study should be smaller than in our replication. But the opposite is the case! This can be explained by a larger trial-by-trial variance in the original study. Larger effect and larger variability in the original study cancel out such that sensitivities are in fact quite comparable to our replication, see below. Further research employing systematic variation of stimulus parameters can further clarify this situation. For example, we are currently determining the role of an ITA in the particular setting of Dehaene et al. (1998) in a more extensive study, see Zerweck et al. (2021).

In summary, we found a similar pattern of results as in the original study: A very poor direct task performance and a clear congruency effect in the indirect task. Based on this pattern of results many researchers would have applied the standard reasoning and inferred an ITA.

Sensitivity Comparison

The appropriate analysis compares sensitivities in direct and indirect tasks. We have already described in the last section that the direct task in our experiment yielded a sensitivity of $d' = 0.26$ ($SD = 0.27$), corresponding to an accuracy of $M = 54.87\%$-correct. For the indirect task, we obtained a sensitivity of $d' = 0.25$ ($SD = 0.15$), corresponding to an accuracy of $M = 54.93\%$-correct ($SD = 3.03$).

Inspection of Figure 2.4a shows that these sensitivities in direct and indirect tasks are very similar, see their difference plot in Figure 2.4d. We found virtually no difference between these sensitivities, $M = -0.01$ ($SD = 0.34$), $t(23) = -0.2$, $p = 0.844$. That is, there is no indication for an ITA.

In conclusion, our results are similar to the typical pattern of results found by Dehaene et al. (1998) and many researchers would have inferred an ITA. However, the appropriate analysis yields no evidence for an ITA: The sensitivities in both tasks are essentially identical.

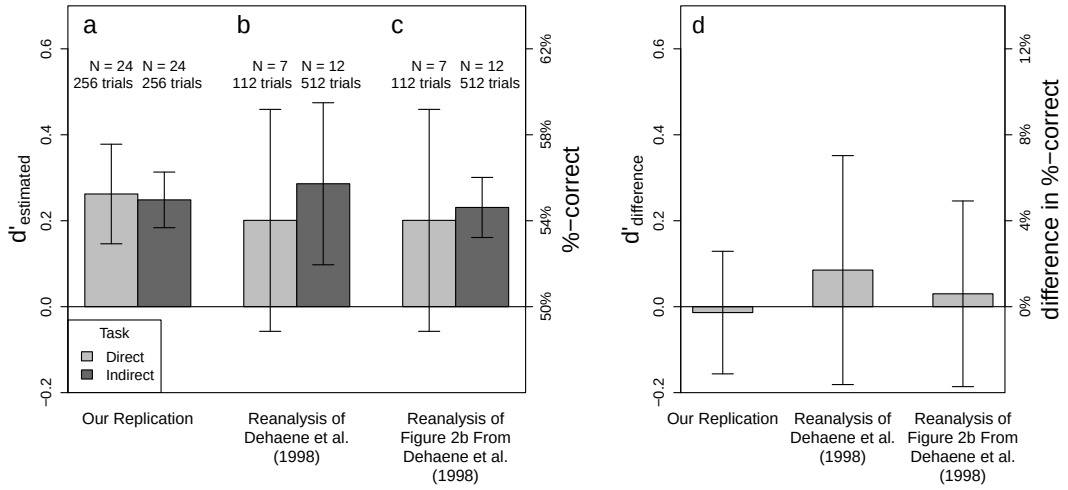

Figure 2.4: **Sensitivities in the Dehaene et al. (1998) setting**. **(a)** Results of our replication study, based on our full trial-by-trial data. **(b)** Results of our reanalysis approach based on the published statistics from Dehaene et al. (1998). **(c)** Reanalysis results from digitizing Figure 2b from Dehaene et al. (1998) showing histograms of indirect task's RT data. For the comparison, we used the same direct task results herein (c) as we used in (b). Comparing (a)–(c) we see that our replication closely matches the results of the original study. **(d)** Difference in sensitivities between direct and indirect tasks: There is no significant difference in sensitivities in our replication study or in our reanalyses of Dehaene et al. (1998). That is, there is no evidence for an ITA. The reanalysis result from (b) is also shown in the large summary in Figure 2.5. Error bars indicate 95% confidence intervals.

2.5 Reanalysis of 15 Influential Studies Finds Hardly any ITA

After having demonstrated that the problems of the widely used standard reasoning are indeed serious, we now apply our approach to a sample of 15 highly relevant studies in the field of unconscious priming.

Methods

Selection Criteria for Reanalyzed Studies

We focused on studies that applied the standard reasoning and claimed an ITA. First, we selected eight studies by hand that are particularly relevant. These studies and their number of citations in Web of Science (Clarivate Analytics, Philadelphia, U.S.A.) are: Finkbeiner and Palermo (2009, 56 citations), Finkbeiner (2011, 13 citations), Mattler (2003, 76 citations), Pessiglione et al.

(2007, 352 citations), Sumner (2008, 34 citations), van Gaal et al. (2010, 154 citations), Y. Wang et al. (2017, 0 citations), Wójcik et al. (2019, 1 citations),

Second, we searched for English articles in Web of Science with the topic "unconscious priming". We selected all studies with more than 150 citations that applied the standard reasoning and claimed an ITA. This resulted in seven additional studies: Damian (2001, 178 citations), Dehaene et al. (1998, 662 citations), Dehaene et al. (2001, 770 citations), Kiefer (2002, 237 citations), Kunde, Kiesel, and Hoffmann (2003, 217 citations), Naccache and Dehaene (2001b, 214 citations), Naccache, Blandin, and Dehaene (2002, 313 citations). Overall, these 15 studies received a total of 3277 citations. See Supplement 2.6.5 for details on these studies.

Details of Analysis When Only Summary Statistics are Available

Our reanalysis method estimates and compares sensitivities for direct and indirect tasks. Here, we sketch some technical details of the analysis. A detailed account with mathematical derivations is given in Supplement 2.6.3.

We denote the estimated sensitivities in the direct and indirect tasks by $d'_{\text{estimated, direct}}$ and $d'_{\text{estimated, indirect}}$, respectively. For the direct task, the typically reported statistics are average d' or %-correct values. Therefore, our estimate is simply the measured sensitivity,

$$d'_{\text{estimated, direct}} = d',$$

or a well-known conversion of %-correct values to d' values assuming neutral observers (Green & Swets, 1988),

$$d'_{\text{estimated, direct}} = 2\Phi^{-1}(\%\text{-correct}),$$

where Φ^{-1} is the inverse of the normal cumulative density function.

In the indirect task, statistics for the congruency effect are typically reported by t values from a paired t test or F values from a repeated–measures ANOVA. In this setting, F values can be translated into t values by $|t| = \sqrt{F}$. From a t value, we estimate the sensitivity by

$$d'_{\text{estimated, indirect}} = t \cdot c_{N,K,q^2} \quad \text{with} \quad c_{N,K,q^2} = \sqrt{\frac{q^2 + \frac{4}{K}}{N}} \sqrt{\frac{2}{N-1}} \frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)}.$$

where Γ is the gamma distribution. The constant c_{N,K,q^2} corrects for the fact that t values increase with increasing number of participants (N), increasing number of trials (K), and that they depend on the ratio of between- and within-subject variance, which we denote by q^2.

The parameter q^2 is the only free parameter we need to estimate for our approach. It is reasonable to assume that this ratio is at most $q^2 = 0.0225$ given our replication study, a literature review (see Supplement 2.6.4) and extensive simulations (see Supplement 2.6.2). Assuming the largest plausible value for q^2, increases the likelihood of finding an ITA thereby following the benefit-of-the-doubt approach.

From the estimated sensitivities, we compute the difference

$$d'_{\text{difference}} = d'_{\text{estimated, indirect}} - d'_{\text{estimated, direct}}$$

and construct a 95% confidence interval using the corresponding standard errors (derived in Supplement 2.6.3). This allows to test for an ITA: If the confidence interval lies above 0 (that is, it has the form $[a, b]$ with $a > 0$), the reported result is significant and an ITA is confirmed, otherwise there is not sufficient evidence to claim an ITA.

We demonstrate in Appendix 2.6.1 that confidence intervals based on our reanalysis method are quite comparable to those based on the trial-by-trial analysis. For the study of ten Brinke et al. (2014), the trial-by-trial analysis versus our reanalysis method using summary statistics produced 95% CI [-0.07; 0.23] and [-0.11; 0.25], respectively (Figure 2.6). In our replication, the two methods produced 95% CI [-0.15; 0.12] and [-0.20; 0.06], respectively (Figure 2.7). Thus, our reanalysis method produces consistent results with the analysis based on trial-by-trial data.

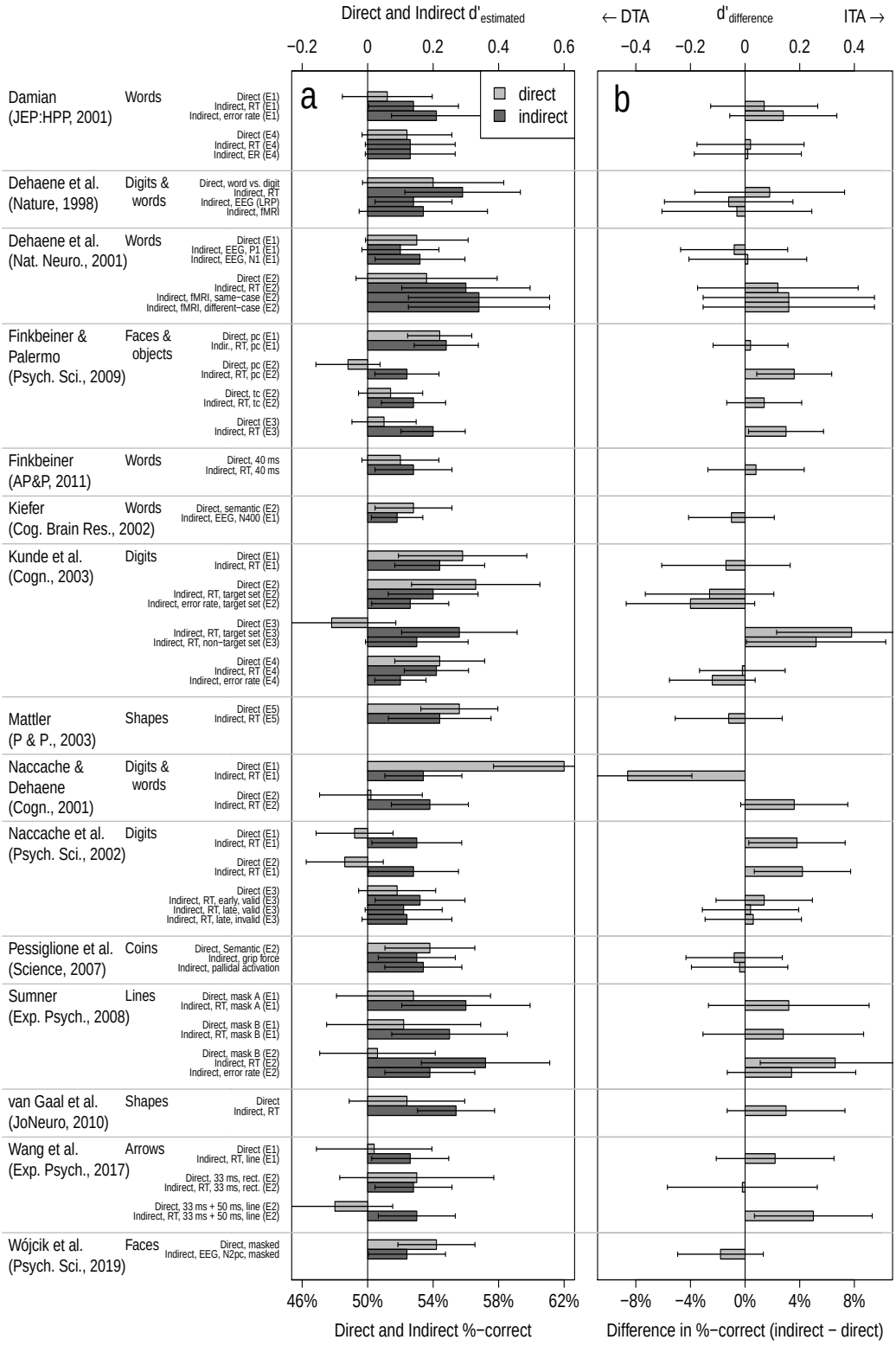
Direct and Indirect d'estimated
← DTA
d'difference
ITA →
a
b
direct
indirect
Damian (JEP:HPP, 2001) Words
Direct (E1)
Indirect, RT (E1)
Indirect, error rate (E1)
Direct (E4)
Indirect, RT (E4)
Indirect, ER (E4)
Dehaene et al. (Nature, 1998) Digits & words
Direct, word vs. digit
Indirect, RT
Indirect, EEG (LRP)
Indirect, fMRI
Dehaene et al. (Nat. Neuro., 2001) Words
Direct (E1)
Indirect, EEG, P1 (E1)
Indirect, EEG, N1 (E1)
Direct (E2)
Indirect, RT (E2)
Indirect, fMRI, same−case (E2)
Indirect, fMRI, different−case (E2)
Finkbeiner & Palermo (Psych. Sci., 2009) Faces & objects
Direct, pc (E1)
Indir., RT, pc (E1)
Direct, pc (E2)
Indirect, RT, pc (E2)
Direct, tc (E2)
Indirect, RT, tc (E2)
Direct (E3)
Indirect, RT (E3)
Finkbeiner (AP&P, 2011) Words
Direct, 40 ms
Indirect, RT, 40 ms
Kiefer (Cog. Brain Res., 2002) Words
Direct, semantic (E2)
Indirect, EEG, N400 (E1)
Kunde et al. (Cogn., 2003) Digits
Direct (E1)
Indirect, RT (E1)
Direct (E2)
Indirect, RT, target set (E2)
Indirect, error rate, target set (E2)
Direct (E3)
Indirect, RT, target set (E3)
Indirect, RT, non−target set (E3)
Direct (E4)
Indirect, RT (E4)
Indirect, error rate (E4)
Mattler (P & P., 2003) Shapes
Direct (E5)
Indirect, RT (E5)
Naccache & Dehaene (Cogn., 2001) Digits & words
Direct (E1)
Indirect, RT (E1)
Direct (E2)
Indirect, RT (E2)
Naccache et al. (Psych. Sci., 2002) Digits
Direct (E1)
Indirect, RT (E1)
Direct (E2)
Indirect, RT (E1)
Direct (E3)
Indirect, RT, early, valid (E3)
Indirect, RT, late, valid (E3)
Indirect, RT, late, invalid (E3)
Pessiglione et al. (Science, 2007) Coins
Direct, Semantic (E2)
Indirect, grip force
Indirect, pallidal activation
Sumner (Exp. Psych., 2008) Lines
Direct, mask A (E1)
Indirect, RT, mask A (E1)
Direct, mask B (E1)
Indirect, RT, mask B (E1)
Direct, mask B (E2)
Indirect, RT (E2)
Indirect, error rate (E2)
van Gaal et al. (JoNeuro, 2010) Shapes
Direct
Indirect, RT
Wang et al. (Exp. Psych., 2017) Arrows
Direct (E1)
Indirect, RT, line (E1)
Direct, 33 ms, rect. (E2)
Indirect, RT, 33 ms, rect. (E2)
Direct, 33 ms + 50 ms, line (E2)
Indirect, RT, 33 ms + 50 ms, line (E2)
Wójcik et al. (Psych. Sci., 2019) Faces
Direct, masked
Indirect, EEG, N2pc, masked
−0.2
0
0.2
0.4
0.6
−0.4
−0.2
0
0.2
0.4
46%
50%
54%
58%
62%
−8%
−4%
0%
4%
8%
Direct and Indirect %−correct
Difference in %−correct (indirect − direct)

Figure 2.5: **Reanalysis of influential studies reporting indirect task advantages (ITAs).** The 15 studies used the standard reasoning to infer an ITA in 44 conditions. **(a)** We reanalyzed the sensitivities and, to foster intuition, we also show %-correct values assuming a neutral observer. **(b)** We reanalyzed the difference in sensitivities: In each group of bars from (a), the indirect task is compared to the corresponding direct task yielding the differences shown in (b). Only if a confidence interval (error bars) around the difference lies to the right and does not contain 0, there is evidence for an ITA. Only in very few cases (8 out of 44), there is evidence for an ITA, while in most cases (35 out of 44) there is no evidence. There is even one case with a significant opposite result, an advantage of the direct task (DTA). Not a single study provides consistent evidence for ITAs across its experiments and conditions in which it claimed ITAs. Moreover, these results are obtained under most favorable conditions for finding an ITA: Our reanalysis overestimates the indirect task sensitivities and therefore the evidence for an ITA due to our conservative choice of analysis strategies. Additionally, some of the reanalyzed studies apply problematic methodology that further biases the results towards finding an ITA even if there is none, see Discussion. This pattern of results casts serious doubts on the existence of ITAs in most, if not all, of the studies. Error bars represent 95%-confidence intervals.

Results and Discussion

We first describe our reanalysis in detail for the study of Dehaene et al. (1998) and then use the same methods for all the other studies.

Reanalysis of Dehaene et al. (1998)

As discussed in our replication, the study reported two direct tasks with sensitivities of $d' = 0.2$ and $d' = 0.3$, respectively. We used the results of the first task, because it had the smaller sensitivity, thereby, increasing the chances of our reanalysis to confirm an ITA and following the benefit-of-the-doubt approach.

In this direct task, $N = 7$ participants were sampled in $K = 112$ trials and a sensitivity of $d' = 0.2$ was reported, see light gray bar in Figure 2.4b. From these values, our reanalysis method estimates the standard error to be $SE = 0.11$.

In the indirect task, the study reported on average a congruency effect of 24 ms with a standard deviation of 13.5 ms in a sample of $N = 12$ participants sampled in $K = 512$ trials each. This equals a t value of $t = 24\text{ ms}/(13.5\text{ ms}/\sqrt{12}) = 6.12$ from which our reanalysis method estimates the sensitivity to be $d'_{\text{estimated, indirect}} = t \cdot c_{N,K,q^2} = 0.29$ ($SE = 0.09$), see dark gray bar in Figure 2.4b.

Taken together, the sensitivities in both tasks are very similar with no clear difference between them, $d'_{\text{difference}} = 0.09$, $SE = 0.14$, see Figure 2.4d. The

confidence interval for the difference includes zero, 95% CI = $[-0.18, 0.35]$, thereby indicating that the sensitivity difference did not deviate significantly from zero. That is, there is no evidence for an ITA.

We were able to reanalyze the results from Dehaene et al. (1998) in an additional way. They depicted summary histograms of RTs in their Figure 2b visualizing that congruent and incongruent RT distributions are similar in shape but only shifted because incongruent RTs were slower than congruent RTs. Despite the shift, RT distributions largely overlap. We digitized the histogram and split RTs along the median as described in the appropriate analysis section. From this, we estimated the indirect task sensitivity to be $d' = 0.23$ ($SE = 0.03$). Again, we find no difference to their first direct task's sensitivity ($d' = 0.2$, $SE = 0.11$) since zero is included in the confidence interval of the difference, 95% CI $[-0.19; 0.25]$, see Figure 2.4c and 2.4d. Note that this approach deviated from our appropriate analysis in that it does not compute the median for each individual participant but uses a grand median across participants because the published histogram pools all participants' RT data. This approach ignores between-subject variance leading to a slight underestimation of the indirect task's sensitivity. Nevertheless, this additional reanalysis provides converging evidence complementing our previous results.

The results from our reanalysis of the original study (Figure 2.4b and 2.4c) and the results from our replication experiment (Figure 2.4a) are very consistent. Estimates for the sensitivities are very stable. This corroborates the validity of our reanalysis approach as well as of our replication experiment (see Supplement 2.6.2 for further validation of our reanalysis approach).

To summarize, both, our reanalysis of Dehaene et al. (1998) as well as our replication of the behavioral responses, suggest that there is no ITA in the behavioral part of that study. This demonstrates the fundamental flaw of the standard reasoning and suggests that similar problems might exist in other studies.

Reanalysis of all 15 studies

We now apply our reanalysis in a similar way to all other studies. For this, we present the data in a more compact fashion in Figure 2.5. For example, what we showed in Figures 2.4b and 2.4d for the study of Dehaene et al. (1998) now corresponds to the lines 7 and 8 in Figure 2.5, showing the sensitivities for each task in Figure 2.5a and the difference of sensitivities in Figure 2.5b.

When evaluating this figure, it is important to be aware that we used our benefit-of-the-doubt approach. For example, Dehaene et al. (1998) had two direct tasks, resulting in $d' = 0.2$ and $d' = 0.3$, respectively. As described above, we used the smaller of those values, thereby increasing the chances of

finding an ITA, which makes our arguments stronger if we nevertheless do not find an ITA (cf. General Discussion).

Inspecting the figure shows that in most studies the sensitivities of direct and indirect tasks have comparable sensitivities, such that the differences are small and not significantly different from zero. This is the case for 35 of the 44 differences between direct and indirect tasks (Figure 2.5b). This is in stark contrast to the fact that all studies claimed ITAs in all these cases.

Only in 8 of the 44 differences there is a significant difference in the direction of an ITA, such that the indirect task has higher sensitivity than the direct task. These results are, however, intermixed with inverted differences in the same studies. For example, although Kunde et al. (2003) have two significant differences in the direction of an ITA, there are five differences pointing in the opposite direction within the same study (albeit those are not significantly different from zero).

Finally, the largest of all differences is even inverted: In Experiment 1 of Naccache and Dehaene (2001b) there is a significantly higher sensitivity in the direct task than in the indirect task, just the opposite of an ITA.

To summarize, our reanalysis found significant ITAs in only 8 out of 44 instances, which are spread across five different studies (Finkbeiner & Palermo, 2009; Kunde et al., 2003; Naccache et al., 2002; Sumner, 2008; Y. Wang et al., 2017). Note that for multiple hypothesis testing, one would expect at least some false positive results. These results are intermixed with 35 inconclusive results and even an opposite result where the direct task showed significantly higher sensitivity than the indirect task (Naccache & Dehaene, 2001b). Inspecting Figure 2.5 shows that there is no consistent evidence for an ITA in any of the reanalyzed studies. Not a single study showed significant ITAs in all conditions, albeit all studies claimed ITAs for all reanalyzed conditions.

Let us stress that our goal was not to investigate whether there exists a "general" ITA across all studies with their vastly different stimuli, experimental setups, tasks and scientific questions. Therefore, we did not perform a meta-analysis or correct for multiple testing. This had several reasons. First, our reanalysis favored finding an ITA by using our benefit-of-the-doubt approach. Second, there are additional methodological issues in the reanalyzed studies that introduce further biases, and for which we cannot correct in our reanalysis (see General Discussion). Considering these two biases towards finding an ITA, a meta-analysis could misleadingly produce the impression that there is a slight ITA present across all reanalyzed studies. An ITA might exist but perhaps only for some particular stimuli and setups. Given that the evidence for an ITA in each individual study is now in question, the research goal should be to differentiate under which conditions a reliable ITA can be obtained and under

which conditions this is not possible. A meta-analysis would not serve this differentiating purpose.

In summary, reanalyzing the results from studies on unconscious priming shows that there is little to no evidence for ITAs in those studies despite them claiming ITAs for all conditions. Sensitivities in the indirect tasks are not consistently larger than sensitivities in the direct task as one would expect, given that unconscious processing was inferred using the standard reasoning that necessarily implies ITAs. This demonstrates how seriously the literature on unconscious priming is affected by the flaws of the standard reasoning.

2.6 Validation and Re-Analysis Details

2.6.1 Validation of Reanalysis Method

We demonstrate in two studies that the appropriate analysis based on the full trial-by-trial data is well approximated by our reanalysis based only on the typically reported statistics (e.g., a t value for the congruency effect in the indirect task). We compare the appropriate analysis using the full, trial-by-trial data on one hand and our reanalysis method based on only the reported summary statistics on the other hand. We apply both approaches to the original data from ten Brinke et al. (2014, see Figure 2.6) and to our replication of Dehaene et al. (1998, see Figure 2.7). Results from the two analyses are very comparable confirming the validity of our reanalysis.

2.6.2 Validation of Reanalysis Method via Simulations

We conducted multiple simulations to validate that our reanalysis method appropriately controls for statistical errors (type I and type II). Each simulation was repeated $10,000$ times. In each run, we generated a trial-by-trial data set with a direct and an indirect task according to the standard repeated measures model outlined in Appendix 2.6.3. We simulated N participants with sensitivities, $d'_{\text{true},i}$, independently and randomly drawn from normal distributions with expected value d'_{true} and variance q^2 (see Appendix 2.6.4 for why q^2 is the variance of individual true sensitivities). Note that we sampled $d'_{\text{true},i}$ for each participant independently in the direct and indirect task to avoid making additional assumptions on their correlation between tasks. Applying Signal Detection Theory, each of these individual sensitivities implies two normal distributions separated by $d'_{\text{true},i}$ standard deviations. From these normal distributions, we sampled a total of K trials for each participant, $K/2$ in each condition. We did this twice, once for each task. In the direct task, we com-

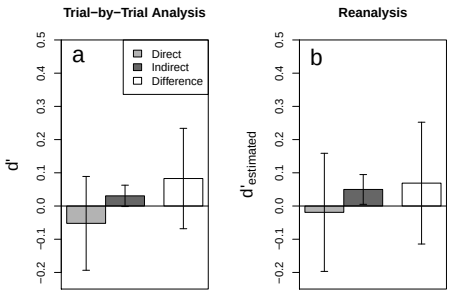

Figure 2.6: **Appropriate analysis applied to ten Brinke et al. (2014)** using the full, trial-by-trial data in (a) and using our reanalysis method in (b). Our reanalysis using only the typically reported statistics produced approximately the same results as the trial-by-trial analysis. In both cases, there is no evidence for an indirect task advantage. Error bars indicate 95% confidence intervals.

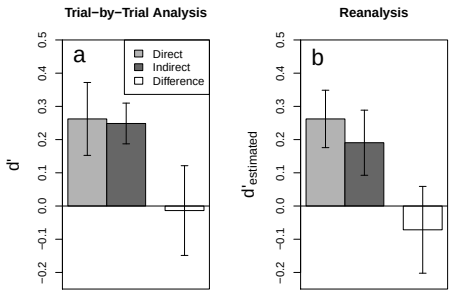

Figure 2.7: **Appropriate analysis applied to our replication of Dehaene et al. (1998)** using the full, trial-by-trial data in (a) and using our reanalysis method in (b). Our reanalysis using only the typically reported statistics produced approximately the same results as the trial-by-trial analysis. In both cases, there is no evidence for an indirect task advantage. Note that the indirect task sensitivity in our reanalysis is smaller than in the trial-by-trial analysis. This is not a contradiction to our claim that *in expectation* the indirect task sensitivity is overestimated by our reanalysis. Estimates of individual studies can vary as indicated by the error bars indicating 95% confidence intervals.

pared each response to the true median: If the response lied on the same side as the normal distribution it was sampled from, the simulated binary decision by the participant in this trial was correct, otherwise it was wrong. In the indirect task, we simply treated the drawn responses as the indirect measures (e.g., RTs). We then applied the traditional analysis used in the standard reasoning and the appropriate analyses, first based on the full, trial-by-trial analysis and second our reanalysis based on typically reported summary statistics. We obtained similar results with log-normal distributions and only report normal distribution results for brevity.

In each simulation, we varied N, K, d'_{true} and q^2. If not declared otherwise, the same q^2 was used for data simulation and reanalysis. Only in simulations 5 and 6, we varied the true q^2 with which the data was simulated and used a different q^2 for our reanalysis in order to see how getting this parameter wrong

would affect our results.

Simulations 1-3 demonstrate that the standard reasoning applied to the traditional analysis miserably fails when applied to the study of Dehaene et al. (1998). Simulation 4 shows that our replication has sufficient statistical power to find an ITA if it was there. Simulations 5 and 6 show how our reanalysis would be affected, if the true q^2 was different than what we assumed. We then summarize additional 108 simulations showing that our estimators, even though they use simplifying approximations, are approximately unbiased.

Simulation 1: Controlling type I errors

We used the same number of participants in the direct ($N = 7$) vs. indirect ($N = 12$) task as well as the same number of trials per condition (direct $K = 112$ vs. indirect $K = 512$) as the original study of Dehaene et al. (1998). Assuming no ITA, we set sensitivities in both tasks to be equal (direct $d'_{\text{true}} = 0.25$ vs. indirect $d'_{\text{true}} = 0.25$). We assumed $q^2 = 0.0225$ for this simulation.

Even though the same sensitivity underlies both tasks, the direct task fails to reach significance half of the time (51.2%) while the indirect task is almost always significant (99.5%). This is not surprising and shows how seriously underpowered the direct task was due to fewer samples, N and K. When applying the standard reasoning, a scientist would erroneously conclude an ITA from a non-significant direct task result and a significant indirect task effect in 48.6% of the experiments. In other words: The widely used standard reasoning would infer an ITA half of the time even though there is no ITA present!

Since there is no ITA present, our reanalysis should find an ITA only as often as prespecified by the significance level $\alpha = 5\%$. Indeed, we find a difference between the two tasks only in 4.7% of the runs. This demonstrates that our reanalysis approach controls appropriately for type I errors.

Simulation 2: Controlling for type II errors with an underpowered direct task

We use the same settings as in Simulation 1 except that we now assume there exists an ITA (direct $d'_{\text{true}} = 0$ vs. indirect $d'_{\text{true}} = 0.25$). Since there is an ITA present, a high statistical power is desired to detect it and avoid type II errors. Typically, a power above $1 - \beta = 80\%$ is desired. However, our reanalysis found the ITA in only 46.2% of the runs. Using the full trial-by-trial data to test for a difference (instead of only using the reported t value from the indirect task) also produced a test power of only 45.9%. There is simply not

enough data in the direct task to provide sufficient evidence for an ITA. The problem with lacking statistical power is not located in our reanalysis because the analysis based on the trial-by-trial data also has a low statistical power. Instead, the problem is the low sample size in the direct task.

Simulation 3: Controlling for type II errors with sufficient samples in the direct task

We repeated Simulation 2 but increased the number of participants and trials in the direct task to match the ones of the indirect task ($N = 12$ and $K = 512$). This is most sensible when testing for a difference because a balanced design maximizes statistical power. Here, our reanalysis method detects the ITA in 78.3% of the runs, which is close to the desired 80%. Using the full trial-by-trial data provides a power of 84.2%. This demonstrates that our reanalysis method provides sufficient power given sufficient samples.

Simulation 4: Statistical power in our replication

We repeated Simulation 3 but used the same number of participants and samples as in our replication study, $N = 24$ and $K = 256$ in both tasks. There, we have the same amount of observations as Dehaene et al. (1998) (double the participants, half the trials). Here, our reanalysis detects the ITA in 96.5% of the runs. The analysis using trial-by-trial data instead of only a t value achieves 97.0%. The increase in statistical power compared to Simulation 3 comes from sampling more participants which is more efficient than sampling more trials given a fixed total number of observations (Rouder & Haaf, 2018).

Simulation 5: Overestimating parameter q^2

We repeated Simulation 3, the balanced Dehaene et al. (1998) setting with an ITA, but generated the data with $q^2 = 0.01$. We still use $q^2 = 0.0225$ for the reanalysis, thus, we overestimate the true q^2. Our reanalysis now successfully detects the ITA in 99.6% of the runs and so does the appropriate analysis with 99.2%. We detect more ITAs here than in Simulation 3 because we make our reanalysis more liberal by choosing a larger q^2.

Simulation 6: Underestimating parameter q^2

Repeating Simulation 5, we now simulated the data with $q^2 = 0.09$ and kept the parameter of our reanalysis at $q^2 = 0.0225$, that is, we now underestimate the true q^2. Individual sensitivities vary a lot now. Even though the mean true direct task sensitivity is $d'_{\text{true}} = 0$ (50%-correct), due to a large standard

deviation of $q = 0.3$, 95% of participants' true sensitivities range between -0.6 (38%-correct) and 0.6 (62%-correct). The assumption $q = 0.3$ poses a problem from a theoretical perspective because some participants can discriminate the masked stimuli relatively well (above 60%-correct). In this case, our reanalysis is more conservative and detects an ITA in only 62.2% of the runs. However, the analysis based on the trial-by-trial data also only achieves a power of 69.2% due to the large variability: Even in this case, our reanalysis would not be too conservative.

Additional Simulations

We conducted additional simulations, one for each combination of the following parameters: $N \in \{5, 10, 20\}$, $K \in \{100, 200, 400\}$, $d'_{\text{true}} = \{0, 0.1, 0.2, 0.5\}$, and $q^2 \in \{0.01, 0.0225, 0.09\}$. In all these simulations, the average, absolute deviation between true and estimated sensitivities was small, $|d'_{\text{true}} - d'_{\text{estimated}}| \leq 0.01$. A deviation of 0.01 in terms of sensitivity translates into a deviation as small as 0.2%-correct, which can be considered negligible in this setting—and deviations in simulations with $N \geq 10$ are substantially smaller.

We computed the standard deviation of $d'_{\text{estimated}}$ (denoted by $SD[d'_{\text{estimated}}]$) across the 10,000 simulations of each parameter combination. We compared this with the estimated standard error, SE. For this purpose, we squared SE of each run, averaged the values and took the square root of the average, which is the standard procedure to average standard errors. For the direct task, the difference between actual variability and our estimates was again $|SD[d'_{\text{estimated}}] - SE| \leq 0.01$. For the indirect task, the same was true when $N \geq 10$. However, for very small sample sizes ($N = 5$) our reanalysis deviated to some degree but the absolute difference between actual standard deviation and our estimates still was $|SD[d'_{\text{estimated}}] - SE| \leq 0.05$. Since all reanalyzed studies use sample sizes of $N \geq 10$ in the indirect task, our reanalysis produced approximately unbiased estimates. Overall, our reanalysis approximates the appropriate analysis sufficiently well in the context we applied it to.

2.6.3 Estimating Sensitivities From Typically Reported Results

We use typically reported results from studies on unconscious priming to estimate the direct and indirect task sensitivities, $d'_{\text{estimated,direct}}$ and $d'_{\text{estimated,indirect}}$. First, we recapitulate the basic model assumptions of a standard repeated measures ANOVA and introduce the notation. We then derive estimators for the sensitivity and standard error in both tasks using only the typically reported

results. Finally, we compute the difference between direct vs. indirect task sensitivities and construct a confidence interval around that difference in order to test for an ITA.

Model assumptions

Our reanalysis of both tasks is based on the standard model of repeated measures ANOVA and paired t test (Winer, Brown, & Michels, 1991; Maxwell & Delaney, 2004; Rouder & Haaf, 2018) as employed in all reanalyzed studies. In this model N participants perform M trials in each condition. In the specific setting we consider, there are only 2 conditions. In the direct task, this corresponds to trials where the masked stimulus is from either of two categories, A vs. B. In the indirect task, the two conditions are typically congruent (A-A, B-B) vs. incongruent (A-B, B-A). In each trial of a given task, Y_{ijk} denotes the response from participant i $(1, ..., N)$ in condition j (1 or 2) in trial k $(1, ..., M)$, where we assume a balanced design such that the total number of trials K is split evenly into the two conditions for $M = K/2$ trials per condition.

In the indirect task, responses Y_{ijk} are the indirect measures (e.g., RTs). In the direct task, it is plausible to assume that responses Y_{ijk} represent participants' internal evidence about the masked stimuli (some neural activity indicating whether the participant saw a masked stimulus from category A or from B). Based on this noisy internal evidence, participants make an internal classification and guess in each trial to which category the stimulus belonged.

The standard model decomposes participants' responses Y_{ijk} into five components:

$$Y_{ijk} = \mu + p_i + c_j + (p \times c)_{ij} + \epsilon_{ijk}.$$

To facilitate understanding, we now describe the model for the example of congruency effects on RTs in the indirect task; but the same notation applies to other indirect measures and to the direct task as well. RTs have a grand mean μ. Some participants have faster RTs than others which is captured in participants' effects p_i. The congruency condition has an effect c_j on RTs. While c_1 is negative leading to faster RTs in congruent trials, c_2 is positive reflecting slower RTs in the incongruent conditions. Participants differ in the extent to which the congruency conditions affect them captured in $(p \times c)_{ij}$ so that some participants have a larger congruency effect than others. The variability in the individual effects is captured by this term's variance, $\text{Var}[(p \times c)_{ij}] = \sigma^2_{p \times c}$. Additionally, there is trial-by-trial noise ϵ_{ijk} from neuromuscular noise and measurement error leading to different responses in each trial. This trial-by-trial measurement error is assumed by the standard models to have a constant variance (homogeneity) across participants and conditions, $\text{Var}[\epsilon_{ijk}] = \sigma^2_\epsilon$. The congruency effect c_j is a fixed effect while participant and interaction

effects (p_i and $(p \times c)_{ij}$) are random effects because they depend on the drawn sample of participants. Random effects and trial-by-trial noise are assumed to be normally distributed with an expected value of zero and their corresponding variance.

Raw effects and sensitivities

Each participant i has an individual expected congruency effect, Δ_i, which theoretically would be obtained by sampling infinitely many trials. The expected RT difference across participants is denoted by Δ.

$$\Delta_i = (c_2 + (p \times c)_{i2}) - (c_1 + (p \times c)_{i1})$$
$$\Delta = c_2 - c_1$$

In a typical experiment, the individual congruency effects are estimated by the observed mean difference between conditions. For the i-th participant, this estimate is $\hat{\Delta}_i$ and averaged across participants this is $\hat{\Delta}$.

$$\hat{\Delta}_i = \bar{Y}_{i2\cdot} - \bar{Y}_{i1\cdot} = \frac{1}{M}\sum_{k=1}^{M} Y_{i2k} - \frac{1}{M}\sum_{k=1}^{M} Y_{i1k}$$
$$\hat{\Delta} = \frac{1}{N}\sum_{i=1}^{N} \hat{\Delta}_i$$

A participant's true sensitivity $d'_{\text{true},i}$ is the normalized effect—normalized by the trial-by-trial error standard deviation σ_ϵ. This quantity indicates, similar to a signal to noise ratio, how well a participant's RTs are separable and therefore to which degree the masked stimuli were processed, cf. Figure 2.3a. The expectation across participants is the true sensitivity d'_{true} indicating how well the RTs of a prototypical participant are separated.

$$d'_{\text{true},i} = \frac{\Delta_i}{\sigma_\epsilon}$$
$$d'_{\text{true}} = \frac{\Delta}{\sigma_\epsilon}.$$

In the direct task, d'_{true} is typically measured by the sensitivity index d' averaged across participants. Participants' individual d'_i are calculated from hit rate, HR (%-correct guesses for masked stimuli from category A), and false alarm rate, FA (%-incorrect guesses for masked stimuli from category B), where

Φ^{-1} is the inverse cumulative density function of the normal distribution.

$$d'_i = \Phi^{-1}(\text{HR}) - \Phi^{-1}(\text{FA})$$
$$d' = \frac{1}{N}\sum_i d'_i.$$

Note that the empirical literature often uses the notation d' without a clear distinction between estimated vs. true value and individual vs. average effects. Because we need to be more precise in our derivations: We denote the true value of an individual participant by $d'_{\text{true},i}$ and the sensitivity index, which is an estimate for the true value, by d'_i. We denote the true sensitivity across participants by d'_{true}. In the direct task, this is estimated by the average across d'_i values denoted by d'. We will also label this averaged estimate $d'_{\text{estimated,indirect}}$.

Two variance sources: true effect (between-) vs. trial-by-trial error (within-subject) variance

Participants differ in their true congruency effect. The variance of these true inter-individual differences can be derived from the model variances using the standard assumptions (1) $(p \times c)_{ij} \sim \mathcal{N}(0, \sigma^2_{p\times c})$, (2) $\text{Var}[c_1] = \text{Var}[c_2] = 0$, and (3) $(p \times c)_{i1} = -(p \times c)_{i2}$. We denote this true effect variance as σ^2_{effect}:

$$\begin{aligned}\sigma^2_{\text{effect}} &= \text{Var}[\Delta_i] = \text{Var}[[c_2 + (p \times c)_{i2}] - [c_1 + (p \times c)_{i1}]] \\ &= \text{Var}[(p \times c)_{i2} - (p \times c)_{i1}] = \text{Var}[2(p \times c)_{i2}] \\ &= 4\sigma^2_{p\times c}.\end{aligned}$$

The variance of the actually observed congruency effects is conceptually different from the variance of the true effects. We denote the variance of the observed congruency effects as $\sigma^2_{\hat{\Delta}_i}$. The observed congruency effects vary more because they are not only affected by true inter-individual difference but also

by trial-by-trial measurement errors:

$$
\begin{aligned}
\sigma^2_{\hat{\Delta}_i} &= \text{Var}[\hat{\Delta}_i] \\
&= \text{Var}\left[\bar{Y}_{i2\cdot} - \bar{Y}_{i1\cdot}\right] \\
&= \text{Var}\left[\frac{1}{M}\left(\sum_{k=1}^{M} \mu + p_i + c_2 + (p \times c)_{i2} + \epsilon_{i2k}\right)\right. \\
&\qquad\qquad \left. - \frac{1}{M}\left(\sum_{k=1}^{M} \mu + p_i + c_1 + (p \times c)_{i1} + \epsilon_{i1k}\right)\right] \\
&= \text{Var}\left[[c_2 + (p \times c)_{i2}] - [c_1 + (p \times c)_{i1}] + \frac{1}{M}\left(\sum_{k=1}^{M} \epsilon_{i2k}\right) - \frac{1}{M}\left(\sum_{k=1}^{M} \epsilon_{i1k}\right)\right] \\
&= \text{Var}\left[\Delta_i + \frac{1}{M}\left(\sum_{k=1}^{M} \epsilon_{i2k}\right) - \frac{1}{M}\left(\sum_{k=1}^{M} \epsilon_{i1k}\right)\right] \\
&= \text{Var}[\Delta_i] + \text{Var}\left[\frac{1}{M}\sum_{k=1}^{M} \epsilon_{i2k}\right] + \text{Var}\left[\frac{1}{M}\sum_{k=1}^{M} \epsilon_{i1k}\right] \\
&= \sigma^2_{\text{effect}} + \frac{2}{M}\sigma^2_\epsilon \\
&= \sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_\epsilon.
\end{aligned}
$$

This has an implication for the variance of average congruency effects, $\hat{\Delta} = \frac{1}{N}\sum_i \hat{\Delta}_i$. These observed, average congruency effects vary due to two variance sources, the true inter-individual differences and trial-by-trial measurement error.

$$\hat{\Delta} \sim \mathcal{N}\left(\Delta, \frac{\sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_\epsilon}{N}\right).$$

We will later have to estimate σ^2_ϵ from a given $\sigma^2_{\hat{\Delta}_i}$. To achieve this, we must disentangle σ^2_{effect} and σ^2_ϵ. We do so by defining the ratio q^2 between these two sources of variance:

$$q^2 = \frac{\sigma^2_{\text{effect}}}{\sigma^2_\epsilon}.$$

This parameter tells us how much of the observed variability comes from true differences vs. noise. If $q^2 = 0$ then all participants would have the same true congruency effect and observed differences are only due to trial-by-trial error. If q^2 is large then there is relatively small trial-by-trial error variance and observed differences between participants stem from reliable, true differences

between participants. Crucially, note that q^2 is also the variance of true, individual sensitivities. Thus, the square root of this ratio, q, is the standard deviation of true, individual sensitivities.

$$Var[d'_{\text{true},i}] = Var\left[\frac{\Delta_i}{\sigma_\epsilon}\right] = \frac{\sigma^2_{\text{effect}}}{\sigma^2_\epsilon} = q^2 \quad \text{corresponding to} \quad SD[d'_{\text{true},i}] = q$$

We derive a reasonable value to use for our setting in Appendix 2.6.4, which is $q^2 = 0.0225$. This means that we will assume that participants' sensitivities $d'_{\text{true},i}$ vary around some true value d'_{true} with a standard deviation of $q = 0.15$.

Relationship between sensitivity and accuracy

As we have already mentioned, some published studies report d' values, whereas other studies report %-correct values in the direct task. Because we would like to be able to work with either of them, we now discuss the relationship that can transform %-correct values into d' values and vice versa.

Recall that $d'_{\text{true},i}$ denotes the true sensitivity of participant i, and let us introduce the notation π_i for the true probability of a correct classification of a masked stimulus by participant i. We now make the assumption of a neutral criterion in the direct task, that is, participants are not inclined to guess one category of the masked stimuli (A or B) more often than the other. Under this assumption, the true relationship is $d'_{\text{true},i} = 2\Phi^{-1}(\pi_i)$ where Φ^{-1} is the inverse cumulative normal distribution (Macmillan & Creelman, 2004). To simplify our later analysis, we now introduce the linear approximation $h(x) = 5(x - 0.5) \approx 2\Phi^{-1}(x)$. This approximation works remarkably well in the regime of sensitivities being close to zero:

$$\text{given } \pi_i, \text{ we approximate } d'_{\text{true},i} \approx h(\pi_i) = 5(\pi_i - 0.5)$$

$$\text{given } d'_{\text{true},i}, \text{ we approximate } \pi_i \approx h^{-1}(d'_{\text{true},i}) = \frac{1}{5}d'_{\text{true},i} + 0.5$$

For example, an accuracy of 54%-correct is approximately translated into the sensitivity $d'_{\text{true},i} \approx 5 \cdot (0.54 - 0.5) = 0.2$. This is very close to the exact translation, $d'_{\text{true},i} = 2\Phi^{-1}(\pi_i) = 0.201$. Table 2.1 shows that this approximation provides a very tight fit in the range of $\pi_i \in [0.4; 0.6]$ or, equivalently, $d'_{\text{true},i} \in [-0.5; 0.5]$. Larger values, that is, an accuracy above 60%-correct, would be at odds with the experimental setting in which direct task performance is assumed to be close to chance ($d'_{\text{true},i}$ close to 0 and π_i close to 0.5).

Table 2.1: Relation between the true accuracy (first column), the approximation of the sensitivity (second column) and the true sensitivity (third column). Note, that for π_i in the range of $[0.5, 0.6]$ and D_i in the range of $[0, 0.5]$ (first six rows in the table) there is a very tight fit between $h(\pi_i)$ and $d'_{\text{true},i}$. Negative values of $d'_{\text{true},i}$ follow symmetrically.

π_i	$h(\pi_i)$	$d'_{\text{true},i}$
0.50	0.000	0.000
0.52	0.100	0.100
0.54	0.200	0.201
0.56	0.300	0.302
0.58	0.400	0.404
0.60	0.500	0.507
0.62	0.600	0.611
0.64	0.700	0.717
0.66	0.800	0.825
0.68	0.900	0.935
0.70	1.000	1.049

Estimated sensitivity, $d'_{\text{estimated,direct}}$, from mean sensitivity index d'

We want to estimate the sensitivity and corresponding standard error from the typically reported direct task results. Usually, the average across individual sensitivity indices is reported as d'. This sensitivity index is already an estimate of the true sensitivity and we take it as it is (Macmillan & Creelman, 2004),

$$d'_{\text{estimated,direct}} = d'. \tag{2.1}$$

The standard error of d' is composed of two variances, one due to systematic variation between individuals' true sensitivities ($d'_{\text{true,i}}$) and the other due to non-systematic measurement error ($\epsilon_{d'_i}$). We use two simplifications: (a) We neglect dependencies between them because the variance of random error $\text{Var}[\epsilon_{d'_i}]$ does not change substantially for different sensitivity values in the relevant range, $D_i^{\text{dir}} \in [-0.5, 0.5]$; (b) We apply the approximation function h that relates d'_i to $\hat{\pi}_i$. This allows us to use the variance of the binomially distributed accuracies $\hat{\pi}_i$ from K trials, $\text{Var}[\epsilon_{\hat{\pi}_i}] = \pi_i(1 - \pi_i)/K$, and relate

them back to the variance of d'_i, which leads to $\mathrm{Var}[\epsilon_{d'_i}] \approx 5^2\,\mathrm{Var}[\epsilon_{\hat{\pi}_i}]$.

$$\begin{aligned} SE_{\text{direct}} &= \sqrt{\mathrm{Var}\left[d'\right]} \\ &= \sqrt{\mathrm{Var}\left[\frac{1}{N}\sum_i d'_i\right]} = \frac{1}{\sqrt{N}}\sqrt{\mathrm{Var}[d'_i]} = \frac{1}{\sqrt{N}}\sqrt{\mathrm{Var}[d'_{\text{true,i}} + \epsilon_{d'_i}]} \\ &\overset{(a)}{\approx} \frac{1}{\sqrt{N}}\sqrt{\mathrm{Var}[d'_{\text{true,i}}] + \mathrm{Var}[\epsilon_{d'_i}]} \overset{(b)}{\approx} \frac{1}{\sqrt{N}}\sqrt{\mathrm{Var}[d'_{\text{true,i}}] + 5^2\,\mathrm{Var}[\epsilon_{\hat{\pi}_i}]} \\ &= \underbrace{\frac{1}{\sqrt{N}}}_{\text{average}} \sqrt{\underbrace{q^2}_{\text{between subject variance}} + 5^2\,\underbrace{\frac{\left(\frac{1}{5}d' + 0.5\right)\left(1 - \left(\frac{1}{5}d' + 0.5\right)\right)}{K}}_{\text{non-systematic error of } \hat{\pi}_i}}. \end{aligned} \tag{2.2}$$

Without simplifications (a) and (b), one could construct an exact estimator. Exact calculations from Miller (1996) show that d' slightly overestimates the true sensitivity d'_{true} but that this bias is so small that the estimator can be considered approximately unbiased when typical sample sizes as in our context are used. On the other hand, our simplifications allow us to find a closed form solution that is simple to compute. Our estimators are well aligned with the true values, which we have shown by validating simulations in Appendix 2.6.2.

Estimated sensitivity, $d'_{\text{estimated,direct}}$, from mean accuracy $\hat{\pi}$

Instead of d', some studies report the average classification accuracy $\hat{\pi}$ (%-correct) for the direct task. We estimate the sensitivity $d'_{\text{estimated,direct}}$ from the mean accuracy $\hat{\pi}$ by a plug-in estimator (Macmillan & Creelman, 2004),

$$d'_{\text{estimated,direct}} = 2\Phi^{-1}(\hat{\pi}) \approx 5 \cdot (\hat{\pi} - 0.5) \tag{2.3}$$

where Φ^{-1} is the inverse cumulative normal distribution. Exploiting the linearity of approximation h in (*), we can derive that this estimator is approximately unbiased:

$$\mathrm{E}[d'_{\text{estimated,direct}}] = \mathrm{E}[2\Phi^{-1}(\hat{\pi})] \approx \mathrm{E}[h(\hat{\pi})] \overset{(*)}{=} h[\mathrm{E}(\hat{\pi})] = h(\pi) \approx 2\Phi^{-1}(\pi) = d'_{\text{true}}.$$

Next, the standard error can be derived in the same fashion as for reported d' values so that we obtain:

$$SE_{\text{direct}} = \frac{1}{\sqrt{N}}\sqrt{q^2 + 5^2\frac{\hat{\pi}\,(1-\hat{\pi})}{K}}. \tag{2.4}$$

Estimated sensitivity, $d'_{\text{estimated,indirect}}$, from t and F values

Now let us move to estimating sensitivities from t values in the indirect task. We will show that an unbiased estimator is obtained from multiplying the t value by the constant c_{N,K,q^2}:

$$d'_{\text{estimated,indirect}} = t \cdot c_{N,K,q^2} \quad \text{with} \quad c_{N,K,q^2} = \sqrt{\frac{q^2 + \frac{4}{K}}{N}} \sqrt{\frac{2}{N-1}} \; \frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)}, \tag{2.5}$$

where Γ is the gamma distribution.

We start by considering how the t value in our setting is computed from the observed congruency effect:

$$t = \frac{\hat{\Delta}}{\hat{\sigma}_{\hat{\Delta}_i}} \sqrt{N}$$

We know that $\hat{\Delta} \sim \mathcal{N}\left(\Delta, (\sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_\epsilon)/N\right)$ from above. Now we introduce independent random variables $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(N-1)$ and rearrange t:

$$\begin{aligned} t &= \frac{Z\sqrt{\left(\sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_\epsilon\right)/N} + \Delta}{\sqrt{\left(\sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_\epsilon\right)}\sqrt{\frac{V}{N-1}}} \sqrt{N} \\ &= \left(Z + \Delta\sqrt{\frac{N}{\left(\sigma^2_{\text{effect}} + \frac{4}{K}\sigma^2_\epsilon\right)}}\right) \frac{\sqrt{N-1}}{\sqrt{V}}. \end{aligned}$$

We now use $\sigma^2_{\text{effect}} = q^2\sigma^2_\epsilon$ (also from above) to isolate σ^2_ϵ and obtain d'_{true}.

$$\begin{aligned} t &= \left(Z + \Delta\sqrt{\frac{N}{\left(q^2\sigma^2_\epsilon + \frac{4}{K}\sigma^2_\epsilon\right)}}\right) \frac{\sqrt{N-1}}{\sqrt{V}} \\ &= \left(Z + \frac{\Delta}{\sigma_\epsilon}\sqrt{\left(\frac{N}{q^2 + \frac{4}{K}}\right)}\right) \frac{\sqrt{N-1}}{\sqrt{V}} \\ &= \left(Z + d'_{true}\sqrt{\left(\frac{N}{q^2 + \frac{4}{K}}\right)}\right) \frac{\sqrt{N-1}}{\sqrt{V}} \end{aligned}$$

As a result, t follows a t distribution with degrees of freedom $df = N-1$ and non-centrality parameter $\delta = d'_{true}\sqrt{\frac{N}{q^2+\frac{4}{K}}}$. From Hogben, Pinkham, and Wilk

(1961) and Hedges (1981), we know the expected t value to be

$$\begin{aligned}\mathrm{E}[t] &= \delta\sqrt{\frac{N-1}{2}}\frac{\Gamma(\frac{N-2}{2})}{\Gamma(\frac{N-1}{2})} \\ &= d'_{true}\cdot\sqrt{\frac{N}{q^2+\frac{4}{K}}}\sqrt{\frac{N-1}{2}}\frac{\Gamma(\frac{N-2}{2})}{\Gamma(\frac{N-1}{2})} \\ &= d'_{\text{true}}\cdot c^{-1}_{N,K,q^2}.\end{aligned}$$

In consequence, an unbiased estimator of d'_{true} is $d'_{\text{estimated, indirect}} = t\cdot c_{N,K,q^2}$.

As for the expected value, the variance of t values is also given by the properties of a non-central t distribution. Multiplying this variance by the constant c_{N,K,q^2} yields the variance of our estimated sensitivity $d'_{\text{estimated, indirect}}$. Since this depends on the non-centrality parameter, we use the plugin estimator

$$\begin{aligned}\hat{\delta} &= d'_{\text{estimated, indirect}}\sqrt{\frac{N}{q^2+\frac{4}{K}}} \\ &= t\cdot c_{N,K,q^2}\sqrt{\frac{N}{q^2+\frac{4}{K}}} \\ &= t\cdot\sqrt{\frac{2}{N-1}}\frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)}\end{aligned}$$

The standard error being its positive square root follows accordingly:

$$\begin{aligned}SE_{\text{direct}} &= \sqrt{\mathrm{Var}[c_{N,K,q^2}\cdot t]} = c_{N,K,q^2}\sqrt{\mathrm{Var}[t]} \\ &= c_{N,K,q^2}\sqrt{\frac{1+\hat{\delta}^2}{N-3}-\frac{\hat{\delta}^2(N-1)\Gamma\left(\frac{N-2}{2}\right)^2}{2\Gamma\left(\frac{N-1}{2}\right)^2}} \\ &= c_{N,K,q^2}\sqrt{\left(1+\frac{2t^2}{N-1}\left(\frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N-2}{2})}\right)^2\right)\left(\frac{N-1}{N-3}\right)-t^2} \qquad (2.6)\end{aligned}$$

With this, we can estimate the sensitivity and its standard error from a given t value in a repeated measures design.

Note that this approach can be applied identically to reported F values instead of t values. The reason is that in repeated measures ANOVA settings with two conditions the equality $|t| = \sqrt{F}$ holds. The main argument can be derived in the following equations using the standard definitions for the

explained (SSE) and residual summed squares (SSR), see Winer et al. (1991); Maxwell and Delaney (2004):

$$t^2 = \left(\frac{\hat{\Delta}}{\hat{\sigma}_{\hat{\Delta}_i}} \cdot \sqrt{N}\right)^2 = \frac{4 \cdot N \cdot (\hat{\Delta}/2)^2}{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\Delta}_i - \hat{\Delta}\right)^2} = \frac{2 \cdot N \cdot (\hat{\Delta}/2)^2}{\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\hat{\Delta}_i - \hat{\Delta}}{2}\right)^2}$$
$$= \frac{2 \cdot N \cdot (\hat{\Delta}/2)^2}{2\sum_{i=1}^{N}\left(\frac{\hat{\Delta}_i - \hat{\Delta}}{2}\right)^2/(2N-2)} = \frac{\mathrm{SSE}/df_{\mathrm{E}}}{\mathrm{SSR}/df_{\mathrm{R}}} = F.$$

Finally, note that this reanalysis for the indirect task can be extended to unbalanced settings in which the total number of trials K is not equally distributed to the two conditions for $M = K/2$ trials per condition but instead to M_1 and M_2 trials per condition $j = 1$ and $j = 2$, respectively. In these situations, one can analogously show that $\hat{\Delta} \sim \mathcal{N}\left(\Delta, (\sigma^2_{\text{effect}} + \frac{M_1+M_2}{M_1 M_2}\sigma^2_\epsilon)/N\right)$. Following the same steps as above, one would obtain an alternative constant that now depends on the split M_1 versus M_2 instead of only K.

$$c_{N,M_1,M_2,q^2} = \sqrt{\frac{q^2 + \frac{M_1+M_2}{M_1 M_2}}{N}}\sqrt{\frac{2}{N-1}}\frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)}.$$

As a sanity check, set $M_1 = M_2 = K/2$ and find $c_{N,M_1,M_2,q^2} = c_{N,K,q^2}$.

Confidence intervals for the difference in sensitivities

Based on the previous estimators, we now need to test for a significant difference between sensitivities in direct vs. indirect tasks. For this purpose we construct a 95% confidence interval around the difference $d'_{\text{difference}}$ while taking the standard error $SE_{\text{difference}}$ of the estimated difference into account:

$$d'_{\text{difference}} = d'_{\text{estimated,indirect}} - d'_{\text{estimated,direct}} \tag{2.7}$$
$$SE_{\text{difference}} = \sqrt{(SE_{\text{direct}})^2 + (SE_{\text{indirect}})^2} \tag{2.8}$$
$$95\%\ \text{CI} = \left[d'_{\text{difference}} \pm z_{0.975} \cdot SE_{\text{difference}}\right], \tag{2.9}$$

where $z_{0.975} = 1.96$ is the 97.5% quantile of the normal distribution. If zero is included in the confidence interval, $0 \in 95\%$ CI, then there is not sufficient evidence for an ITA because the observed difference can be explained by measurement error in a situation where the true direct and indirect task sensitivities are equal. Only if the confidence interval lies above zero, that is 95% CI $= [a, b]$ and $a > 0$, there is evidence for the presence of an ITA.

Note that in this test we use quantiles z_α of the normal distribution and not quantiles of the t distribution. Using the t distribution would require to estimate the degrees of freedom, which is unnecessarily complicated for our approach. We use the quantiles of the normal distribution which leads to a more liberal test increasing the likelihood of confirming an ITA and following the benefit-of-the-doubt approach (see General Discussion).

2.6.4 Estimating the Ratio q^2 of Between- vs. Within-Subject Variance

As we have seen in the reanalysis of direct and indirect task sensitivities, we need to know one parameter: q^2, a ratio of systematic vs. noise variance. This is not an artifact of our reanalysis but unavoidable.

What Does the Parameter q^2 Mean?

To see what this parameter means and why we need to estimate it, consider estimating the indirect task sensitivity $d'_{\text{estimated,indirect}}$ from t values. A t value is computed by dividing an observed effect by its standard error, $t = \bar{x}/SE$. In the indirect task, $\bar{x}$ may be the average congruency effect and SE the estimated standard error of congruency effects across participants. This standard error is influenced by two sources of variability: variance due to inter-individual differences in true congruency effects across participants (σ^2_{effect}) and variance due to trial-by-trial measurement error (σ^2_ϵ). We want to isolate the latter variance, σ^2_ϵ, because we want to estimate the underlying sensitivity $d'_{\text{true}} = \Delta/\sigma_\epsilon$ from the t value. Thus, we need to distinguish the two sources of variability. We do so by defining the ratio q^2:

$$q^2 = \frac{\sigma^2_{\text{effect}}}{\sigma^2_\epsilon}.$$

Note that this parameter is equal to the variance of individual true sensitivities, $q^2 = Var[d'_{\text{true},i}]$, see Supplement 2.6.3. Therefore, it might be more intuitive to consider the un-squared parameter, which is the standard deviation of participants' true sensitivities, $q = \text{SD}[d'_{\text{true},i}]$.

Literature Review to Determine q^2 (Following Benefit-Of-The-Doubt Approach)

To estimate q^2, we consider multiple studies that either provide estimates or make explicit assumptions. All these studies yield a specific value, see our summary in Table 2.2, columns q^2 and q. For our reanalysis, we will use the largest plausible value, $q^2 = 0.0225$. Thus, we follow the benefit-of-the-doubt

approach giving a previously established ITA the best chance to be confirmed in our reanalysis.

Table 2.2: We repeated our reanalysis of the indirect task sensitivity from Dehaene et al. (1998) (last column) based on the q^2 values from different studies. Larger values of q^2 increase the estimated, indirect task sensitivity. We took the largest plausible value our reanalysis method.

Study	q^2	q	Reanalysis of Dehaene et al. (1998) $d'_{\text{estimated, indirect}}$
ten Brinke et al. (2014)	**0.0020**	0.04	0.16
Our example study	**0.0074**	0.09	0.20
Rouder & Haaf (2018)	**0.0087**	0.09	0.21
Miller & Ulrich (2013)	**0.0121**	0.11	0.23
Jensen (2002)	**0.0142**	0.12	0.25
Ribeiro et al. (2016)	**0.0214**	0.15	0.28
Our assumption	**0.0225**	0.15	0.29

First, we estimated q^2 from the data of ten Brinke et al. (2014). This yielded $\hat{\sigma}^2_{\text{effect}} = (6.5 \text{ ms})^2$ and $\hat{\sigma}^2_{\epsilon} = (144 \text{ ms})^2$, which translates into an estimated ratio of $\hat{q}^2 = (6.5 \text{ ms})^2/(144 \text{ ms})^2 = 0.0020$.

Our replication based on Dehaene et al. (1998) produced estimates for the variances of $\hat{\sigma}^2_{\text{effect}} = (6.7 \text{ ms})^2$ and $\hat{\sigma}^2_{\epsilon} = (78 \text{ ms})^2$ translating into an estimated ratio of $\hat{q}^2 = 0.0074$.

Similarly, Rouder and Haaf (2018, p. 21) discuss the relation between the two sources of variance in psychophysics. Their formulas are identical to ours when changing the notation from σ^2_{effect} to σ^2_{β} and σ^2_{ϵ} to σ^2. They argue that reasonable values are $\sigma_{\text{effect}} = 28$ ms and $\sigma_{\epsilon} = 300$, which leads to $q^2 = \sigma^2_{\text{effect}}/\sigma^2_{\epsilon} = 0.0087$.

Other studies did not discuss the ratio between the two variances, σ^2_{effect} and σ^2_{ϵ}, but only the trial-by-trial error variability σ^2_{ϵ}. We can combine this with Dehaene et al. (1998) reporting the observed standard deviation of RT effects to be 13.5 ms. This variability is constituted by $\hat{\sigma}^2_{\hat{\Delta}_i} = \hat{\sigma}^2_{\text{effect}} + \frac{4}{K}\hat{\sigma}^2_{\epsilon} = (13.5 \text{ ms})^2$. By knowing $\hat{\sigma}^2_{\epsilon}$ and the number of trials, K, we can rearrange the formula and estimate $\hat{\sigma}^2_{\text{effect}}$ and thereupon $\hat{q}^2$.

Miller and Ulrich (2013, p. 846, in their Table 3) suggested $\sigma_{\epsilon} = 96$ ms in a binary forced-choice task (without masked stimuli): Their error term E_k with variance $Var[E_k] = 91.5$ corresponds to the mean noise across 100 trials, see their Table 15. From this, we obtained $\sigma_{\epsilon} = \sqrt{Var[E_k] \cdot 100} = 96$ ms, as noted above. Combining this with Dehaene et al.'s results yields $\sigma_{\text{effect}} = 10.5$ ms and thereupon $q^2 = 0.0121$.

Jensen (1992, p. 877, Table 7 for task "Hick SS 2") reported an average estimate of $\hat{\sigma}_\epsilon = 91$ ms measured in $N = 863$ nine to twelve year olds yielding $q^2 = 0.014$. Ribeiro, Paiva, and Castelo-Branco (2016) report $\hat{\sigma}_\epsilon = 79$ ms in a speeded binary choice task without priming suggesting $q^2 = 0.021$. Even though the specific tasks and populations from these last two studies do not match Dehaene et al.'s setting exactly, it is plausible that variances are by and large comparable.

Given this range of parameter values, we use a value that is larger than any q^2 value reported in these studies:

$$q^2 = 0.0225 \quad \text{corresponding to} \quad q = SD[d'_{\text{true},i}] = 0.15.$$

By choosing this upper bound on q^2, we follow the benefit-of-the-doubt approach because large values of q^2 favor the ITA hypothesis in our reanalysis attributing more variance to σ^2_{effect} and less to σ^2_ϵ. This, in turn, increases our estimate of $d'_{\text{true}} = \Delta/\sigma_\epsilon$. For example, see how larger values of q^2 increase the estimated indirect task sensitivity from Dehaene et al. (1998) in the last column of Table 2.2. Hence, overestimating q^2 leads to an overestimation of the indirect task sensitivity increasing the chances of confirming an ITA.

How Would our Reanalysis Look Like With a Different q^2?

We have repeated our literature reanalysis from Figure 5 with different parameter values. In Figure 2.8, we show a more realistic reanalysis with $q^2 = 0.01$. Here, the picture resembles a null effect. In contrast, we show an overly optimistic reanalysis with $q^2 = 0.09$ in Figure 2.9, in which an ITA starts to emerge for many studies. However, even in this case there is no conclusive evidence for an ITA in most studies because confidence intervals for the sensitivity difference still include 0.

Note that, when assuming large q^2 values as in Figure 2.5 and 2.9, one cannot take the reanalysis result as *evidence for* an ITA. This is because large q^2-values bias our reanalysis in favor of finding an ITA. Only when we nevertheless do not find an ITA, these results can be meaningfully interpreted as *evidence against* an ITA. In order to establish evidence for an ITA, one would have to use smaller values for q^2 or, better yet, use the trial-by-trial data so that no assumption on q^2 is necessary. Otherwise, an apparent ITA result may only be due to the bias introduced by a large q^2.

We provide an online tool to perform our reanalysis with different values of q^2 at `http://www.ecogsci.cs.uni-tuebingen.de/ITAcalculator/`. There, we suggest three different values for q^2: To establish a lack of evidence for an ITA, we suggest $q^2 = 0.0225$ as in our reanalysis proper. This assumption rarely rejects evidence for an ITA if there is any. On the other hand, to

establish evidence for an ITA, we instead suggest $q^2 = 0.0025$, which is more restrictive. Only when an ITA is established with a relatively small q^2 like this, we can be sure that it is a genuine ITA instead of being produced only due to a lenient assumption on q^2. Lastly, we suggest an intermediate value of $q^2 = 0.01$, which is suitable for an exploratory reanalysis. Note that depending on the exact experimental setup, different values of q^2 may be appropriate.

Overall Summary Regarding our Choice of q^2

Taken together, our replication, our simulations, and the literature review suggests that q^2 is clearly below 0.0225. We adopted this upper bound as our assumption because it increases the chances of finding an ITA, thereby, following the benefit-of-the-doubt approach. We use this assumption to show that evidence for an ITA is missing in many studies. To establish evidence for an ITA, the reanalysis would have to use smaller values to rule out the possibility that an ITA was only the product of the overestimation bias coming from a too large q^2.

Up to now, we have only discussed behavioral data (RTs) but we applied our reanalysis method also to EEG and fMRI data. The justification for this is that the relative noise level is even larger in single-trial event related potentials (ERPs) and blood-oxygen-level-dependent signals (BOLD signals) because they incorporate much more noise (Stahl, Gibbons, & Miller, 2010). Thus, the ratio between effect vs. noise variance in these measures will be even smaller, again, justifying our choice of $q^2 = 0.0225$.

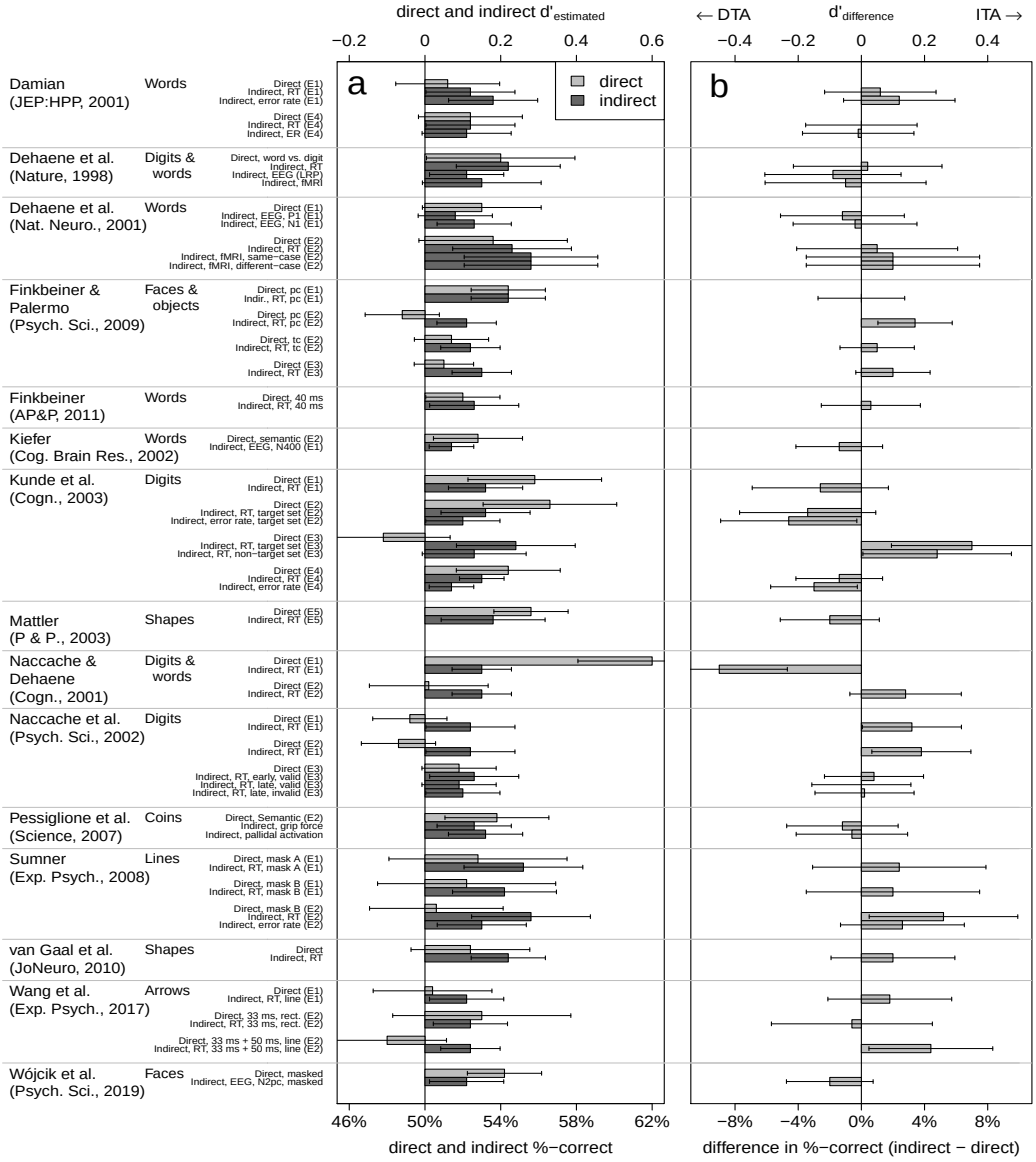

Figure 2.8: **Reanalysis with $q^2 = 0.01$.** Same as Figure 2.5 assuming that the standard deviation of true sensitivities across participants is $\mathrm{SD}[d'_{\mathrm{true},i}] = q = 0.1$. This assumption matches the results of our replication and is therefore more realistic but also more strict in dismissing results of an indirect task advantage (ITA). Here, only 7 reanalyzed ITAs are confirmed while 3 results yield the opposite result of a larger sensitivity in the direct task (direct task advantage [DTA]). Error bars represent 95%-confidence intervals.

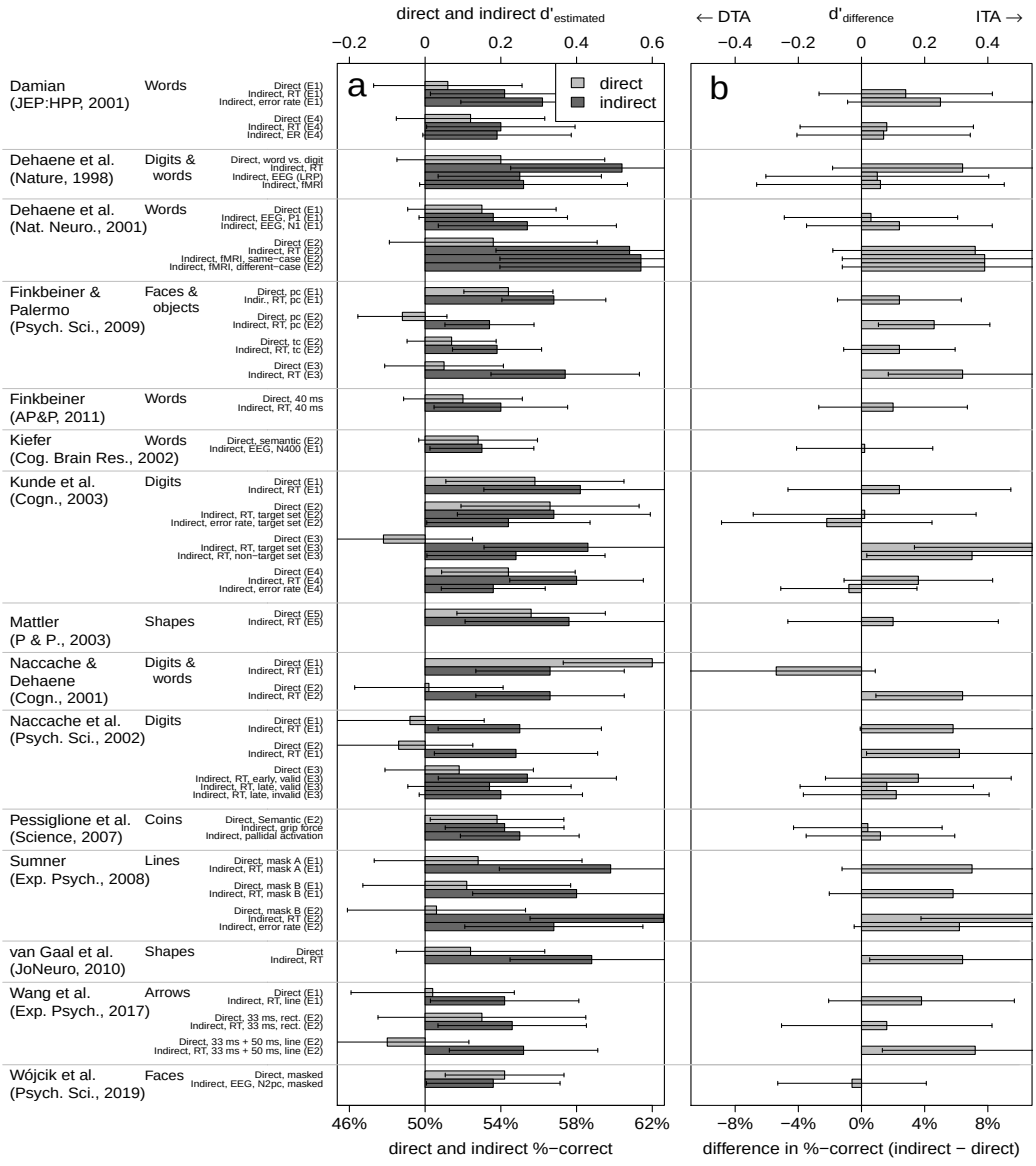

Figure 2.9: **Reanalysis with $q^2 = 0.09$.** Same as Figure 2.5 assuming that the standard deviation of true sensitivities across participants is $\mathrm{SD}[d'_{\mathrm{true},i}] = q = 0.3$. With this or even larger q^2, reanalyzed sensitivities tend to become clearly larger in the indirect compared to the direct task. However, this assumption is clearly unrealistic. First, in the direct task, this would mean that a substantial percentage of participants had a true sensitivity of $d'_{\mathrm{true},i} = 0.5$ or higher indicating that they could discriminate the masked stimuli better than 60%-correct. In the indirect task, an unrealistic implication of this assumption is that, in the study of Dehaene et al. (1998), trial-by-trial reaction times (RTs) would be estimated to vary with a standard deviation of only ±43 ms (within-subject variance $\sigma^2_\epsilon = 43^2$) even though RTs typically vary more than ±80 ms from trial to trial, see Appendix 2.6.4.

2.6.5 Details of Reanalyzed Studies

For each study, we give an overview of the study's structure, indicate in a table which values we extracted and explain our decisions for in- and exclusion of particular results. We only use results that follow the standard reasoning, claim an ITA and fit into our reanalysis method. We include quotes from the reanalyzed studies indicating their adherence to the standard reasoning. We use the following two abbreviations:

NR **N**ot **r**eanalyzable: Reported statistics do not match our reanalysis method. For example when the congruency factor has more than two levels (congruent, incongruent, and neutral) or when there are additional between-subject factors.

NIE **N**o **i**ndirect **e**ffect: The study attempted to find an ITA but failed due to a non-significant indirect task result. In such cases, the studies usually abort the standard reasoning, such that these cases are not relevant for us.

We report the number N of participants, the total number of trials K, and the reported statistic of the original study. Additionally, we report the sensitivities and standard errors according to our reanalysis. These are the values from Figure 2.5a. We then report the differences in sensitivities and their standard errors; here the difference is always taken between the current row's indirect task compared to the previously reported direct task. These results are presented in Figure 2.5b. We abbreviate Experiment 1 by E1, etc.

We also mark studies that excluded participants with good direct-task performance by adding the label **Regression to the mean** (see Discussion on why this is problematic). We still reanalyzed the reported results, although the exclusion introduced a bias for which our reanalysis method does not correct. This bias is liberal and favors finding an ITA. Thus, we follow the benefit-of-the-doubt approach.

Damian (2001)

The study reports four experiments but concludes an ITA only in Experiment 1 and 4. Experiments 2 and 3 were NIE.

Standard Reasoning: "Two control experiments investigated participants' ability to consciously perceive the masked primes. It was shown that performance was at chance level on both presence-absence judgments and on a number vs. random letter string discrimination task when the temporal characteristics of a trial were identical to those of the main experiment. Thus, the congruity effect described above must indeed have occurred outside of the participants' awareness" (p. 1).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	16	96	$d' = 0.064$	0.06 ± 0.07	
Indirect, RT (E1)	16	120	$F(1, 15) = 6.15$	0.14 ± 0.07	0.07 ± 0.10
Indirect, error rate (E1)	16	120	$F(1, 15) = 13.8$	0.21 ± 0.07	0.14 ± 0.10
Direct (E4)	16	96	$d' = 0.117$	0.12 ± 0.07	
Indirect, RT (E4)	16	120	$F(1, 15) = 5.67$	0.13 ± 0.07	0.02 ± 0.10
Indirect, ER (E4)	16	120	$F(1, 15) = 5$	0.13 ± 0.07	0.01 ± 0.10

Dehaene et al. (1998)

The study reported two direct tasks and three indirect tasks. From the two direct tasks, we consider only the second direct task (word vs. digit discrimination) because it fits the neutral criterion assumption and it also shows lower sensitivity ($d' = 0.2$ in the first and $d' = 0.3$ in the second task). This way, we favor confirming the ITA hypothesis. For the first indirect measure, we computed the t value from the given estimates for the congruency effect ($M = 24$ ms and $SD = 13.5$). For the second indirect measure, the statistic ($t(11) < 3$) is taken from Figure 4, where the covert activation reflects processing of the prime as opposed processing of the target in the overt activation. For the third indirect measure, we only considered the congruency effect on fMRI the results are provided in Figure 5.

Standard Reasoning: "Under these conditions, even when subjects focused their attention on the prime, they could neither reliably report its presence or absence nor discriminate it from a nonsense string (Table 1). Nevertheless, we show here that the prime is processed to a high cognitive level [by demonstrating a priming effect]."

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, word vs. digit	7	112	$d' = 0.2$	0.20 ± 0.11	
Indirect, RT	12	512	$t(11) = 6.16$	0.29 ± 0.09	0.09 ± 0.14
Indirect, EEG (LRP)	12	512	$t(11) = 3$	0.14 ± 0.06	-0.06 ± 0.12
Indirect, fMRI	9	128	$F(1, 8) = 6.23$	0.17 ± 0.10	-0.03 ± 0.14

Dehaene et al. (2001)

The study reports two experiments. In E1, multiple measures assessed the visibility of the masked stimulus and we chose the reported binary forced-choice task (no stimulus vs. masked stimulus) because it is the most relevant result. In this experiment, the ITA refers to the absence vs. presence of the masked stimuli. The fMRI results in E1 were NR. In E2, the ITA referred to the congruency effect of repeated (congruent, either in same or in different case) vs. different words (incongruent).

Standard Reasoning: "Behaviorally, participants again denied seeing the primes and were unable to select them in a two-alternative forced-choice test [...]. However, case-independent repetition priming was observed in response times recorded during imaging [...]" (p. 755) and "As this phenomenon depends only on the identity of the masked prime, specific information about word identity must have been extracted and encoded unconsciously [...]" (p. 756).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	27	36	52.9%-correct	0.15 ± 0.09	
Indirect, EEG, P1 (E1)	12	300	$t(11) = 2.04$	0.10 ± 0.06	-0.04 ± 0.10
Indirect, EEG, N1 (E1)	12	300	$F(1, 11) = 9.79$	0.16 ± 0.07	0.01 ± 0.11
Direct (E2)	10	64	53.6%-correct	0.18 ± 0.11	
Indirect, RT (E2)	10	480	$F(1, 9) = 36$	0.30 ± 0.10	0.12 ± 0.15
Indirect, fMRI, same-case (E2)	10	240	$t(9) = 1.98$	0.34 ± 0.11	0.16 ± 0.16
Indirect, fMRI, different-case (E2)	10	240	$t(9) = 2.68$	0.34 ± 0.11	0.16 ± 0.16

Finkbeiner and Palermo (2009)

The study reported four experiments. Prime and target stimuli were presented in different locations to the participants. In half of the trials the prime location was cued (pc) and in the other half it was the target location (tc). We excluded the target cued condition in E1 because it was NIE. In E3, multiple within-subject factors were tested but since those do not change the reported F value of the congruency effect we could nevertheless reanalyze it. E4 did not follow the standard reasoning.

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, pc (E1)	40	80	$d' = 0.22$	0.22 ± 0.05	
Indirect, RT, pc (E1)	40	80	$F(1, 39) = 33.94$	0.24 ± 0.05	0.02 ± 0.07
Direct, pc (E2)	40	80	$d' = -0.06$	-0.06 ± 0.05	
Indirect, RT, pc (E2)	40	80	$F(1, 39) = 8.5$	0.12 ± 0.05	0.18 ± 0.07
Direct, tc (E2)	40	80	$d' = 0.07$	0.07 ± 0.05	
Indirect, RT, tc (E2)	40	80	$F(1, 39) = 10.6$	0.14 ± 0.05	0.07 ± 0.07
Direct (E3)	20	240	$d' = 0.05$	0.05 ± 0.05	
Indirect, RT (E3)	20	720	$F(1, 19) = 31.37$	0.20 ± 0.05	0.15 ± 0.07

Finkbeiner (2011, Regression to the mean)

The study presented trials in two conditions, one with a short (40 ms) and one with a long (50 ms) prime presentation duration. An ITA was concluded only for the short duration and with respect to the semantic content (not color).

Standard Reasoning: "In contrast, 16 of the 21 subjects were judged to be at chance with the 40-ms primes. Following Rouder et al. (2007), the RTs for the 17 subject-by-prime-duration combinations for which subliminality was confirmed were entered into a paired-samples t test (two-tailed) to determine whether subliminal priming had occurred" (p. 1260).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, 40 ms	21	120	$d' = 0.098$	0.10 ± 0.06	
Indirect, RT, 40 ms	21	80	$t(20) = 2.5$	0.14 ± 0.06	0.04 ± 0.09

Kiefer (2002)

The study reported two experiments. E1 reported the indirect task results and E2 reported the direct task results. In E1, indirect effects on RT, error rates and some EEG components were NR because the reported statistics combine masked and unmasked conditions (for unmasked conditions, they claimed no ITA) except for the N400 component in EEG. In E2, there were multiple direct tasks (see their Table 1). We chose the direct task on semantic judgment because the indirect task's congruency effect was an effect from semantic relatedness too.

Standard Reasoning: "Average d' measures in all tasks and context conditions did not deviate significantly from zero demonstrating that masked words were not identified" (p. 36).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, semantic (E2)	24	80	$d' = 0.14$	0.14 ± 0.06	
Indirect, EEG, N400 (E1)	24	320	$F(1, 23) = 5.48$	0.09 ± 0.04	-0.05 ± 0.08

Kunde et al. (2003)

The study reported four experiments. In E1, there were multiple direct task measures from which we chose the one that fit our model assumptions of a neutral criterion (the identification rate is not comparable by our method). Also in E1, we chose not to consider sub-analyses of the indirect effects because they are essentially repetitions of the same comparison. In E2, we did not consider the non-target set condition and in E3 we did not consider the error

rate analysis as they were NIE. In E1-E3, trials with neutral primes were not considered for calculating the priming effect.

Standard Reasoning: "The identification rate for the prime numbers was 2.2% (the chance level is 6.25% as each prime is presented four times in the 64 test trials). Thus, the primes were indeed unidentifiable, as is usually found under the experimental conditions that we adopted (Damian, 2001; Dehaene et al., 1998; Koechlin et al., 1999; Naccache & Dehaene, 2001)" (p. 230).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	12	64	$d' = 0.29$	0.29 ± 0.10	
Indirect, RT (E1)	12	1152	$F(1, 11) = 25.17$	0.22 ± 0.07	-0.07 ± 0.12
Direct (E2)	12	64	$d' = 0.33$	0.33 ± 0.10	
Indirect, RT, target set (E2)	12	288	$F(1, 11) = 15.24$	0.20 ± 0.07	-0.13 ± 0.12
Indirect, error rate, target set (E2)	12	288	$F(1, 11) = 6.35$	0.13 ± 0.06	-0.20 ± 0.12
Direct (E3)	12	64	$d' = -0.11$	-0.11 ± 0.10	
Indirect, RT, target set (E3)	12	144	$F(1, 11) = 21.67$	0.28 ± 0.09	0.39 ± 0.14
Indirect, RT, non-target set (E3)	12	144	$F(1, 11) = 6.58$	0.15 ± 0.08	0.26 ± 0.13
Direct (E4)	24	64	$d' = 0.22$	0.22 ± 0.07	
Indirect, RT (E4)	24	1152	$F(1, 23) = 43.2$	0.21 ± 0.05	-0.01 ± 0.08
Indirect, error rate (E4)	24	1152	$F(1, 23) = 9.17$	0.10 ± 0.04	-0.12 ± 0.08

Mattler (2003)

The study reports five experiments. Only Experiments 3 and 5 are considered to be evidence for unconscious priming. Experiment 3 suffers severely from regression to the mean and is therefore not reanalyzed.

Standard Reasoning: "We might assume that performance at chance level indexes absence of all conscious information. This assumption was made in a number of studies (e.g., Dehaene et al., 1998; Klotz & Neumann, 1999; Neumann & Klotz, 1994; Vorberg et al., in press). In the present study, evidence for priming without awareness comes from Experiment 3 and Experiment 5, in which participants showed substantial non-motor priming effects although they could not discriminate primes better than chance" (p. 184)

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E5)	11	320	$d' = 0.28$	0.28 ± 0.06	
Indirect, RT (E5)	11	320	$F(1, 10) = 18.5$	0.22 ± 0.08	-0.06 ± 0.10

Naccache and Dehaene (2001b)

The study reports two experiments. For the direct tasks in both experiments, the authors additionally conducted the Greenwald method (Greenwald, Draine, & Abrams, 1996; Draine & Greenwald, 1998) which, however, has been

criticized before (Dosher, 1998; Klauer, Greenwald, & Draine, 1998; Miller, 2000; Merikle & Reingold, 1998). Therefore, we only considered typical results as in all other studies. We considered only the main congruency effects on RT and no further subanalyses because the reported direct task would not have been comparable. In both experiments, an old and a new stimulus set were used. In E1, we only reanalyzed the RT effect based on the old stimulus set because the direct task sensitivity was estimated only for the old set. In E2, we reanalyzed the RT effect for the mixed, both new and old, stimulus set because the direct task sensitivity was estimated for this mixed set, too.

Standard Reasoning: "In this task, subjects performed at chance level, while priming effects were replicated. This study provides strong evidence for the unconscious nature of our semantic priming effects" (p. 227).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	18	32	$d' = 0.6$	0.60 ± 0.11	
Indirect, RT (E1)	18	384	$F(1, 17) = 21.99$	0.19 ± 0.06	-0.41 ± 0.12
Direct (E2)	18	64	$d' = 0.01$	0.01 ± 0.08	
Indirect, RT (E2)	18	384	$F(1, 17) = 21.62$	0.19 ± 0.06	0.18 ± 0.10

Naccache et al. (2002)

The study reported three experiments. We did not consider the subanalyses for cued trials as the standard reasoning only related to the congruency effects. Note that we only counted the number of "critical" trials which were used in their analysis.

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	12	240	$d' = -0.04$	-0.04 ± 0.06	
Indirect, RT (E1)	12	240	$F(1, 11) = 7.88$	0.15 ± 0.07	0.19 ± 0.09
Direct (E2)	12	240	$d' = -0.07$	-0.07 ± 0.06	
Indirect, RT (E1)	12	240	$F(1, 11) = 7.32$	0.14 ± 0.07	0.21 ± 0.09
Direct (E3)	12	240	$d' = 0.09$	0.09 ± 0.06	
Indirect, RT, early, valid (E3)	12	240	$F(1, 11) = 9.23$	0.16 ± 0.07	0.07 ± 0.09
Indirect, RT, late, valid (E3)	12	240	$F(1, 11) = 3.97$	0.11 ± 0.06	0.02 ± 0.09
Indirect, RT, late, invalid (E3)	12	240	$F(1, 11) = 5.34$	0.12 ± 0.07	0.03 ± 0.09

Pessiglione et al. (2007, Regression to the mean)

The study deviated from the standard priming paradigm by just showing masked stimuli (in this case, coins) and no target stimuli. Presentation duration was varied in three conditions. For the separate conditions, participants were measured in one direct task and with three indirect measures. The appendix provided the required information for our reanalysis. We digitized their

Figure S2 to derive the t values for the two indirect measures grip force and pallidal activation. The third indirect measure, skin conductance, was NIE. Even though these results were only reported in the appendix, the study bases their interpretation on these results. Note, that $N = 24$ relates to 24 participant $\times$ stimulus duration conditions in which the direct task was non-significant at an individual level.

Standard Reasoning: "Based on the percentage of correct responses, the analysis could then be restricted to all situations where subjects guess at chance level about stimulus identity (fig. S2) [by removing situations with significant direct task results]. Even in these situations, pallidal activation and handgrip force were significantly higher for pounds as compared to pennies [...] " (p. 906).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, Semantic (E2)	24	60	$d' = 0.19$	0.19 ± 0.07	
Indirect, grip force	24	90	$t(23) = 2.92$	0.15 ± 0.06	-0.04 ± 0.09
Indirect, pallidal activation	24	90	$t(23) = 3.41$	0.17 ± 0.06	-0.02 ± 0.09

Sumner (2008, Regression to the mean)

The study reported two experiments. Both, E1 and E2, had different mask conditions (A vs. B). Only E1 provided indirect task results such that we could reanalyze both conditions separately. For E2 we had to apply our reanalysis to both conditions aggregated. Therefore, we averaged over the given d' values from both conditions. We did not consider the subanalyses on the difference and interaction between the two masks but only the congruency effects as they are taken for the standard reasoning.

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, mask A (E1)	12	40	$d' = 0.14$	0.14 ± 0.12	
Indirect, RT, mask A (E1)	12	200	$t(11) = 5.5$	0.30 ± 0.10	0.16 ± 0.15
Direct, mask B (E1)	12	40	$d' = 0.11$	0.11 ± 0.12	
Indirect, RT, mask B (E1)	12	200	$t(11) = 4.5$	0.25 ± 0.09	0.14 ± 0.15
Direct, mask B (E2)	12	80	50.5%-correct	0.03 ± 0.09	
Indirect, RT (E2)	12	400	$t(11) = 7.4$	0.36 ± 0.10	0.33 ± 0.14
Indirect, error rate (E2)	12	400	$t(11) = 4$	0.19 ± 0.07	0.17 ± 0.12

van Gaal et al. (2010, Regression to the mean)

The study reported one experiment with one direct task and multiple indirect measures. However, we only considered the indirect effect on RTs as the fMRI analyses were NR.

Standard Reasoning: "[...] a, Participants were unable to discriminate between trials with a strongly masked square or diamond, as revealed by chance-level performance in a two-choice discrimination task administered after the main experiment. b, Although strongly masked no-go signals could not be perceived consciously, they still triggered inhibitory control processes, as revealed by significantly longer response times on these trials than on strongly masked go trials." (in Figure 2, p. 4145).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct	20	48	$d' = 0.118$	0.12 ± 0.09	
Indirect, RT	20	240	$t(19) = 6.24$	0.27 ± 0.06	0.15 ± 0.11

Y. Wang et al. (2017)

The study reported two experiments. In E1, there were two outline conditions, line vs. rectangle. The line condition yielded a negative congruency effect which we treated similar to a standard (positive) priming effect. The rectangle condition was NIE. In E2, the rectangle condition with prime duration of 50 ms produced a large d' so that no ITA was claimed. Hence, we only considered the rectangle condition only for 33 ms. For the line condition, 33 ms and 50 ms trials were analyzed together since there was no interaction effect.

Standard Reasoning: "The results from the FC task indicated that similar prime visibility, equivalent to chance level, was obtained in the two preposed object type conditions. This finding confirmed that primes were processed subliminally in the primary task" (p. 425).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct (E1)	15	64	$d' = 0.02$	0.02 ± 0.09	
Indirect, RT, line (E1)	15	208	$F(1, 14) = 6.86$	0.13 ± 0.06	0.11 ± 0.11
Direct, 33 ms, rect. (E2)	15	32	$d' = 0.15$	0.15 ± 0.12	
Indirect, RT, 33 ms, rect. (E2)	15	208	$F(1, 14) = 8.15$	0.14 ± 0.06	-0.01 ± 0.14
Direct, 33 ms + 50 ms, line (E2)	15	64	$d' = -0.103$	-0.10 ± 0.09	
Indirect, RT, 33 ms + 50 ms, line (E2)	15	416	$F(1, 14) = 11.47$	0.15 ± 0.06	0.25 ± 0.11

Wójcik et al. (2019)

The study reported one experiment with masked and unmasked conditions. We only considered the masked condition for which an ITA was claimed but not the unmasked condition. In the direct task, we had to compute average d' from the openly accessible material. In the indirect task, EEG components were measured. For EEG preprocessing, some trials had to be rejected leading

to an average of 131 trials. We assumed that rejection rate was approximately equal in the two indirect task conditions.

Standard Reasoning: "Analysis of the sensitivity measure d' indicated that faces were not consciously identified in the masked condition. A clear N2 posterior-contralateral (N2pc) component (a neural marker of attention shifts) was found in both the masked and unmasked conditions, revealing that one's own face automatically captures attention when processed unconsciously" (in the abstract, p. 471).

	Original data			Our reanalysis (Figure 5)	
	N	K	Statistic	$d'_{\text{estimated}} \pm SE$	$d'_{\text{diff}} \pm SE_{\text{diff}}$
Direct, masked	18	160	$d' = 0.211$	0.21 ± 0.06	
Indirect, EEG, N2pc, masked	18	131	$t(17) = 2.34$	0.12 ± 0.06	-0.09 ± 0.08

2.6.6 Cost of Dichotomization in Significance Testing and Bayesian Analyses

The main fallacy of the standard reasoning persists independently of which statistical methods are chosen (significance testing or Bayesian analysis). It comes from evaluating the two tasks separately instead of using the appropriate analysis of measuring a difference between direct vs. indirect task sensitivities. To see why problems occur in both methods, consider the following simulation demonstrating the cost of dichotomization.

In multiple runs, we simulate one data set by sampling responses from $N = 12$ participants and $K = 256$ trials per participant. Thus, we sample $K/2 = 128$ observations in each of two conditions based on two normal distributions that are shifted by $d'_{\text{true}} = 0.15$ standard deviations (corresponding to a true performance of 53%-correct; log-normal distributions produce similar results). We analyze this one data set (a) as in the indirect task and (b) as in the direct task. We will show that both methods, significance testing and Bayesian analysis, produce misleading results in favor of the indirect task even though the *exact same data* is the basis for both tasks.

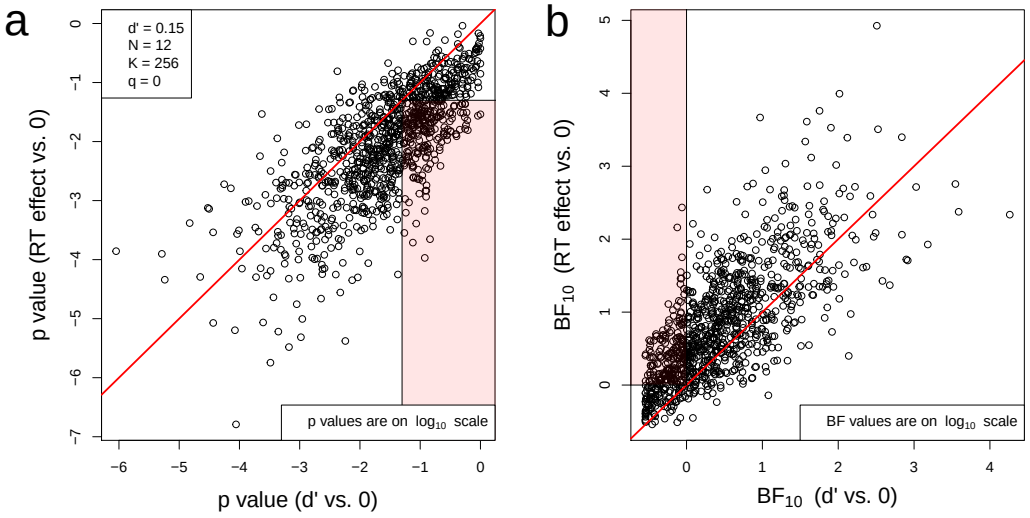

Figure 2.10: **Cost of Dichotomization in significance testing and Bayesian analysis.** Each point corresponds to one simulated data set. We analyzed each data set as in the direct task (x axis) and indirect task (y axes). We find that p values in (a) as well as Bayes Factors in (b) diverge from the red equality line indicating more evidence in the indirect task due to the loss of information from median splitting the data in the direct task. Shaded regions indicate a misleading pattern of result: (a) a significant indirect task vs. a non-significant direct task result; (b) a Bayes Factor supporting the null hypothesis in the direct task vs. a Bayes Factor supporting the alternative hypothesis in the indirect task.

To mimic the RT effect from the indirect task, we tested the mean difference between two conditions against 0 (y axes in Figure 2.10). To mimic the direct task, we conducted a median split and tested sensitivity d' against 0 (x axes in Figure 2.10). The assumption here is that participants have access to the same information in both tasks and were forced to give a binary response (dichotomize) in the direct task so that the best they could do is to respond according to the optimal median split criterion. To test against 0, we used a t test (see Figure 2.10a) and we computed Bayes Factor (Figure 2.10b) using the R package provided by Morey, Rouder, Jamil, and Morey (2015)

Inspecting the results in Figure 2.10, we find that p values and Bayes Factors diverge from the red equality line indicating more evidence for an effect in the indirect task analysis compared to the direct task analysis. This is so because a median split dichotomization discards information (Cohen, 1983) producing larger p values and smaller Bayes Factors in the direct as compared to the indirect task.

In 23% of the simulations, there is a non-significant direct task vs. a significant indirect task result (shaded area in Figure 2.10a). This pattern may mislead researchers into thinking that there is an effect in the indirect task but none in the direct task. Note that this is a well–known error: One cannot take a non-significant result as evidence for the absence of an effect without a power analysis (see for example Vadillo et al., 2020).

The pattern of results from Bayes Factors is misleading in an even more severe way. In 20% of the simulations, we find Bayes Factors supporting the null hypothesis of no effect in the direct task ($BF_{10} < 1$) and simultaneously supporting the alternative hypothesis in the indirect task ($BF_{10} > 1$; on the log scale these are values below and above 0, see shaded area in Figure 2.10b). We even found some simulations, in which there is substantial evidence for the *null* hypothesis in the direct task ($BF_{10} < 1/3$) and substantial evidence for the *alternative* hypothesis in the indirect task ($BF_{10} > 3$). That is, if we ignored the main fallacy of the standard reasoning and followed the Bayesian analysis naively, we would conclude a difference in the two tasks even though the analyses in both tasks is based on the exact same data!

Analyzing the simulated data separately—computing mean difference in the indirect task and sensitivity in the direct task—produces misleading patterns of results. This problem occurs independent of the statistical methods used, significance testing or Bayes analysis, and even if the exact same data underlies both tasks. In a real experiment, direct and indirect tasks would not be based on the exact same data but on two samples, which produces additional measurement error. But in our idealized simulation here, there is no additional sampling error because both tasks are based on the same sample.

Hence, no difference between the two tasks should be found. Accordingly, the appropriate analysis based on the sensitivity comparison would find exactly $d'_{\text{estimated, indirect}} - d'_{\text{estimated, direct}} = 0$ correctly identifying no difference between the two tasks and solving this problem.

2.7 General Discussion

Many studies on consciousness that investigate a wide range of cognitive functions are based on the flawed standard reasoning. The main fallacy occurs when the standard reasoning infers an ITA. That is, a higher sensitivity for masked stimuli in the indirect task as compared to the direct task. In an earlier reanalysis of ten Brinke et al. (2014) by Franz and von Luxburg (2015), in our replication of the behavioral part of Dehaene et al. (1998), and in our reanalysis of 15 highly influential studies, we found that none of these studies can overall truly claim evidence for an ITA. To the contrary, responses in the indirect task often show a similar sensitivity as compared to the direct task. This casts serious doubt on the evidence for unconscious processing that exceeds conscious reportability in these studies.

The fallacy of the standard reasoning has serious consequences for the trustworthiness of the scientific literature on consciousness. It also takes away from the appeal of many claims in the field: For example, it would be an interesting result if lie detection and semantic meaning of numbers could be processed outside of awareness. But such strong claims require substantive empirical evidence, which we did not find because the reanalyzed studies employed the flawed standard reasoning. The appropriate analysis yields results that may be considered as less exciting because—under scrutiny—participants' responses did not seem to be affected by processing beyond what they can consciously report.

Besides theoretical issues, there are also additional methodological problems that can systematically bias the results and lead to claims of an ITA even if the true underlying sensitivities in the direct and the indirect task are perfectly equal.

First, a common practice is to exclude participants with a good direct task sensitivity. The researchers' motivation here is to avoid including the subset of participants who are consciously aware of the masked stimuli. However, this practice bears the problem of regression to the mean (Barnett, van der Pols, & Dobson, 2004; Schmidt, 2015; Shanks, 2017). Thus, this practice is biased towards finding a smaller sensitivity in the direct task and thus biased towards finding an ITA even if there is none. Several studies in our reanalysis have this problem (Finkbeiner, 2011; Mattler, 2003; Pessiglione et al., 2007; Sumner,

2008; van Gaal et al., 2010). This can explain why these studies produced some of the largest differences in our reanalysis in Figure 2.5.

Second, in some experimental procedures participants have to respond to the target stimulus (indirect task) first and only then respond to the masked stimulus (direct task) all within the same trial (see Finkbeiner & Palermo, 2009; Peremen & Lamy, 2014). Because the cognitive impact of a masked stimulus decays quickly after 300 ms (Mattler, 2005; Wolfe, 1999), this procedure makes the direct task more difficult. Participants have to memorize the masked stimulus while performing the indirect task until they can give a direct task response. This may decrease the direct task sensitivity due to the additional difficulty, which can produce misleading ITA results. It is somewhat impressive that, even under these favorable circumstances, none of these reanalyzed studies provide consistent evidence for an ITA.

Nevertheless, our results do not necessarily rule out the possibility that ITAs exist in some cases. But the existence of an ITA may depend on the particular task and stimuli used. It might not be as ubiquitous as previously thought. Albeit the long standing request to use the same metric for both tasks (Reingold & Merikle, 1988) has often been ignored, there are some studies that provide evidence for an ITA using the appropriate analysis. For example, the setting of Schmidt (2002)—color stimuli served as primes and targets—found a distinct ITA result. Another example is the study by Kunst-Wilson and Zajonc (1980) using geometric shapes (but see also de Zilva, Vu, Newell, & Pearson, 2013; Seamon, Brody, & Kauff, 1983).

Therefore, we do not claim that there are no instances in which an ITA exists. Such a claim would be far beyond the scope of a single scientific study. But we do claim that one of the most prevalent methods in the wide research area of unconscious priming is fundamentally flawed. This flaw affects and potentially invalidates interpretations of many studies. As a consequence, the field has to reassess the situation of ITAs by applying the appropriate analysis to substantiate or refute previously made claims.

In deriving our appropriate methods, we have chosen strategies that favored finding an ITA. That is, we have followed the benefit-of-the-doubt approach to increase the chances of confirming an ITA. From such an approach, one would have expected clear evidence for an ITA in each of the reanalyzed studies. But since we nevertheless did not find consistent evidence for ITAs, having followed the benefit-of-the-doubt approach makes our arguments even stronger.

However, in future research, we hope that the benefit-of-the-doubt approach will no longer be necessary because it has a drawback: It would be inappropriate to simply revert the reasoning and use our liberal method to establish evidence *for* an ITA. To provide convincing evidence for an ITA, we

would need a more balanced approach, one that might have not convinced researchers in the current situation (because they might have rejected it for being too conservative in terms of finding an ITA). For example, we used a clearly fail-safe estimate for q^2 in our reanalysis, that was chosen to be *larger* than all reported values on which this estimate is based. A more balanced approach would use a smaller estimate, which would reduce the chances to find an ITA, see our additional reanalyses in Supplement 2.6.4 for a figure like Figure 2.5 but with a more balanced estimate of q^2. Of course, trial-by-trial data should be used whenever possible.

To summarize, what we suggest is a research program: Given the tremendous interest in unconscious priming and the far-reaching inferences based on studies using the standard reasoning, researchers should reinvestigate the most relevant cases of claimed ITAs and clarify to which degree the claims in those studies are truly warranted. In those cases where an ITA is properly established, researchers can then start to draw further reaching conclusions about conscious vs. unconscious processing (Eriksen, 1960; Erdelyi, 1986; Holender, 1986; Reingold & Merikle, 1988; Schmidt & Vorberg, 2006). An ITA is only a prerequisite but not a sufficient condition for the inferences that are typically drawn about unconscious processing.

In short, the literature needs a serious and concerted reassessment that would go well beyond the scope of a single study and will also require—in critical cases—the collection of new data. In many cases where superior unconscious processing already seemed an established fact (e.g., Hassin, 2013), we expect that this view needs to be revised. In other cases, researchers might still be able to establish such a relationship—which will then be even more interesting and foster the theoretical understanding of when exactly conscious processing is vital for a cognitive function and when it is not.

2.8 Glossary for Chapter 2

Variable	Description
c_j	Condition effect, for example the congruent condition ($j = 1$) produces faster RTs so that $c_1 < 0$ and $c_2 = 1 - c_1 > 0$.
c_{N,K,q^2}	Constant relating t values to the estimated sensitivity in the indirect task, $d'_{\text{true}} = c_{N,K,q^2} \cdot t$. It depends on N, M and q.
d'	Observed, average sensitivity index, estimates the true sensitivity d'_{true}.
d'_i	Observed, individual sensitivity indices, estimates the true, individual sensitivities $d'_{\text{true},i}$.
d'_{true}	True sensitivity, $d'_{\text{true}} = \frac{\Delta}{\sigma_\epsilon}$.
$d'_{\text{estimated}}$	Estimated sensitivity from the reported summary statistics in the direct ($d'_{\text{estimated,direct}}$) or indirect task ($d'_{\text{estimated,indirect}}$).
$d'_{\text{true},i}$	Individual sensitivity, $d'_{\text{true},i} = \frac{\Delta_i}{\sigma_\epsilon}$.
Δ	The true difference between conditions, $\Delta = c_2 - c_1$.
$\hat{\Delta}$	The observed, mean difference between conditions.
Δ_i	True, individual effects, $\Delta_i = c_2 + (p \times c)_{i2} - (c_1 + (p \times c)_{i1})$, for example the expected congruency effect between conditions of participant i.
$\hat{\Delta}_i$	The observed difference between conditions of participant i.
ϵ_{ijk}	trial-by-trial error, noise due to measurement error or random neuronal fluctuations.
$f_{opt}(x)$	Optimal classifier taking indirect measures x (e.g., RTs) and predicting the condition (congruent/incongruent).
$f_t(x)$	Threshold classifier predicting one condition for indirect measures $x \leq t$ (e.g., RTs) and the other for $x > t$.
h	Linear approximation used to translate between sensitivities and accuracies.
i	Index for participant $i \in \{1, 2, ..., N\}$.
j	Index for condition $j \in \{1, 2\}$, for example indicator for congruent ($j = 1$) and incongruent ($j = 2$) conditions.
K	Total number of trials per particpant, $K = 2M$.
k	Index for trial $k \in \{1, 2, ..., M\}$. Since there are two conditions, the number of observed trials per participant is $2M = K$.

Variable	Description
M	Number of trials per participant × condition. The total number of trials per participant is $2M = K$.
μ	Grand mean, for example the overall expected value of RTs.
N	Number of participants.
$\Omega(x)$	Marginal, cumulative density distribution (CDF) over indirect measures x.
p_i	Participant effect, for example participants with a faster RTs than average have a negative p_i while slower participants have a positive p_i.
$(p \times c)_{ij}$	Interaction effect, for example some participants have different reaction time effects.
π	True accuracy.
$\hat{\pi}$	Observed, mean accuracy.
π_i	True accuracy of participant i. It can be translated into a sensitivity by $d'_{\text{true},i} = 2\Phi^{-1}(\pi_i)$ where Φ is the cumulative normal distribution.
$\hat{\pi}_i$	Observed, individual accuracy.
q^2	Ratio between effect variance and trial-by-trial error variance, $q^2 = \frac{\sigma^2_{\text{effect}}}{\sigma^2_\epsilon}$. This is the variance of true sensitivities across individuals, $q^2 = \text{Var}[d'_{\text{true},i}]$. A reasonable value in our setting is $q^2 = 0.0225$ implying $\text{SD}[d'_{\text{true},i}] = 0.15$.
SE	Estimated standard error of the estimated sensitivity.
$\sigma^2_{\Delta_i}$	Variance of true individual effects, for example, to which degree participants vary in their congruency effect.
$\sigma^2_{\hat{\Delta}_i}$	True variance of observed individual effects, for example, variance of the observable congruency effects.
$\hat{\sigma}^2_{\hat{\Delta}_i}$	Estimated variance of observed individual effects. This is what scientists get when computing the variance on the observable congruency effects across participants.
$\sigma^2_{p\times c}$	Variance of the interaction effect, $(p \times c)_{ij}$.
σ^2_{effect}	Variance of the effects Δ_i, $\sigma^2_{\text{effect}} = 4\sigma^2_{p\times c}$.
σ^2_ϵ	Variance of the trial-by-trial error, ϵ_{ijk}.
t	t value, in our context it comes from paired-t-tests between the two conditions of the indirect task.
Y_{ijk}	Response of participant i in condition j trial k from the direct (Y^{dir}_{ijk}) or indirect task (Y^{indir}_{ijk}). The standard repeated measures ANOVA model is $Y_{ijk} = \mu + p_i + c_j + (p\times c)_{ij} + \epsilon_{ijk}$.

Chapter 3

Predicting Group Decisions Based on Confidence Weighted Majority Voting

Published as: Meyen, S., Sigg, D. M. B., von Luxburg, U., & Franz, V. H. (2021). Group Decisions Based on Confidence Weighted Majority Voting. *Cognitive Research: Principles and Implications*. https://doi.org/10.1186/s41235-021-00279-0

Abstract

Background: It has repeatedly been reported that, when making decisions under uncertainty, groups outperform individuals. Real groups are often replaced by simulated groups: Instead of performing an actual group discussion, individual responses are aggregated by a numerical computation. While studies have typically used unweighted majority voting (MV) for this aggregation, the theoretically optimal method is confidence weighted majority voting (CWMV)—if independent and accurate confidence ratings from the individual group members are available. To determine which simulations (MV vs. CWMV) reflect real group processes better, we applied formal cognitive modeling and compared simulated group responses to real group responses.

Results: Simulated group decisions based on CWMV matched the accuracy of real group decisions, while simulated group decisions based on MV showed lower accuracy. CWMV predicted the confidence that

groups put into their group decisions well. However, real groups treated individual votes to some extent more equally weighted than suggested by CWMV. Additionally, real groups tend to put lower confidence into their decisions compared to CWMV simulations.

Conclusion: Our results highlight the importance of taking individual confidences into account when simulating group decisions: We found that real groups can aggregate individual confidences in a way that matches statistical aggregations given by CWMV to some extent. This implies that research using simulated group decisions should use CWMV instead of MV as a benchmark to compare real groups to.

3.1 Significance Statement

The question of how a group determines an overall group decision from the individual votes of its group members is pervasive and likely as old as mankind. It is at the basis of democratic voting rules and is also prevalent with new urgency in the age of the Internet, where often many individual votes, or ratings, are available that one wants to combine to an optimal overall group decision—without there being the possibility of real group discussions. From a theoretical point of view, the situation is clear: Individual confidences should be taken into account and confidence weighted majority voting (CWMV) is the statistically optimal aggregation procedure (under quite general assumptions). However, in research on group decisions, CWMV is not routinely used for comparison to real group performances, but instead the simpler majority vote (MV) that ignores the individual confidences. Therefore, it is currently not clear whether real groups weigh individual votes in the same way CWMV does. Real groups may be limited in their capacity to take individual confidence ratings into consideration or may rely on different strategies. We compared real group decision to simulated group decisions based on the CWMV and MV procedures. We found that real groups weigh individual confidences in a way that can be well described by CWMV. These results suggest that basic research as well as online-based aggregation of individual votes or ratings could benefit from using CWMV instead of MV.

3.2 Background

Under uncertainty, groups make more accurate decisions than individuals (Koriat, 2015; Mannes et al., 2014): Medical students achieve more accurate diagnoses in groups than individually (Hautz et al., 2015); medical diagnoses improve when groups of independent doctors are involved (Kurvers et al., 2016; Wolf, Krause, Carney, Bogart, & Kurvers, 2015); groups of students make more accurate judgments about criminal cases than individuals (van Dijk et al., 2014); groups detect lies more accurately than individuals (Klein & Epley, 2015); groups achieve higher IQ scores than individuals (referred to as wisdom of the crowd, Bachrach, Graepel, Kasneci, Kosinski, & Van Gael, 2012; Vercammen, Ji, & Burgman, 2019; Kosinski et al., 2012) etc. Exceptions occur when group members have widely different levels of competence (Galesic, Barkoczi, & Katsikopoulos, 2018; Puncochar & Fox, 2004; van Dijk et al., 2014). Nevertheless, groups generally outperform individuals.

Although some of the above mentioned studies also used real groups (Hautz et al., 2015; Klein & Epley, 2015; van Dijk et al., 2014), all of these studies simulated group decisions: Individuals gave responses that were then statistically aggregated into one simulated group response without a real group discussion occurring. A crucial aspect is therefore the choice of aggregation method that is used to simulate group decisions. One frequently used method is majority voting (MV; Hastie & Kameda, 2005; and see for example Klein & Epley, 2015; van Dijk et al., 2014; Kosinski et al., 2012; Kurvers et al., 2016; Sorkin et al., 2001).

In MV, the most frequent individual decision (vote) is taken as the simulated group decision. By design, MV weighs all individual responses equally. Note, however, that real groups typically perform better than simulated groups using MV (Bahrami et al., 2010; Birnbaum & Diecidue, 2015; Klein & Epley, 2015; Sniezek & Henry, 1989). This shows that MV cannot capture all the processes that are at work in real group decisions.

In particular, MV overlooks that individuals can estimate how accurate their own decisions are in many situations (Brenner et al., 1996; Fleming et al., 2012; Griffin & Tversky, 1992; Martins, 2006; Zehetleitner & Rausch, 2013; Regenwetter et al., 2014) even though there are also situations in which they cannot (Klein & Epley, 2015; Koriat, 2012b, 2017; Litvinova, Kurvers, Hertwig, & Herzog, 2019). When reliable confidence estimates are available, they can influence real group discussions: It is plausible that individuals share their sense of confidence during group interactions (Bang et al., 2014) such that votes from confident individuals are weighted more than those of less confident individuals.

There are methods that have taken confidence ratings from individuals into account. One of the most prominent is the maximum-confidence slating algorithm by Koriat (2012a, 2012b). In this algorithm, the most confident individual decides the vote. Another approach for dealing with multiple confidence ratings is to not only consider the most confident individual but a small subgroup of the top most confident individuals (Mannes et al., 2014), or to average all confidences (Litvinova et al., 2020). However, these methods to simulate group decisions do not strictly follow the mathematically optimal way to aggregate confidences.

The theoretically optimal method to aggregate individual confidences is confidence weighted majority voting (CWMV; Grofman et al., 1983; Nitzan & Paroush, 1982)—assuming that individuals can accurately assess confidences in their independently formed decisions. CWMV aggregates these independent responses (votes and confidences) in the mathematically optimal way by giving more weight to reliable than unreliable votes. Thus, statistically aggregating individual responses into a simulated group decision using CWMV rather than MV may reflect real groups better and provide a more appropriate benchmark.

Do real, interacting groups weigh individual confidences in a way that is reflected by simulating a group discussion using CWMV? It is not clear whether real group decisions are adequately represented by CWMV, since CWMV is only sporadically applied in current research. Bahrami et al. (2010) found that group performance of dyads is well predicted by CWMV. Hautz et al. (2015) found that real dyads performed better than CWMV, which predicts the group response of a dyad to be that of the most confident member. CWMV is also discussed in animals from an evolutionary perspective (Marshall, Brown, & Radford, 2017). However, to our knowledge, no study has yet considered groups with more than two members comparing decisions from real group discussions versus simulated decisions using CWMV on a trial-by-trial basis.

In our experiment, we investigated whether CWMV simulations can predict real group decision of triads (groups of three). We compared simulated group decisions to real group decisions on a trial-by-trial basis. Our groups consisted of three individuals because we wanted to investigate whether real groups weigh confidences in a way that is adequately reflected by CWMV. In contrast, using only dyads, CWMV simulates the group decisions to be the vote of the more confident individual (similar to maximum-confidence slating) and CWMV can only contribute by predicting a dyad's combined confidence based on the individual responses. But triads can display qualitatively different behavior than dyads: While it is sometimes the case that the most confident individual determines the group decision in triads, triads also allow for the possibility that the most confident individual is overruled by the two other

group members when they are sufficiently confident in the alternative choice. Thus, we want to clarify whether real groups of three weigh individual votes in a way that can be characterized by CWMV.

Before describing our experiment, we will give a more formal description of the simulation methods MV and CWMV. We will present a formal cognitive model (e.g., see Forstmann, Wagenmakers, Eichele, Brown, & Serences, 2011) that allows us to measure in how far real groups deviate from CWMV simulations.

3.2.1 Majority voting (MV) versus confidence weighted majority voting (CWMV)

CWMV assumes that multiple individuals report independent decisions (votes) as well as confidence ratings. These confidence ratings indicate how reliable individual decisions are. CWMV weighs the decisions by the confidence ratings in a theoretically optimal way to simulate a group decision (Grofman et al., 1983; Nitzan & Paroush, 1982). This section shortly introduces the basic mathematical notation, first of MV and then of CWMV.

Let a group consist of n individuals. The task is to decide between multiple (usually two) options from which exactly one is correct. For example, consider $n = 3$ students trying to determine whether a suspect of a criminal case is guilty or not (cf. van Dijk et al., 2014). First, each individual forms a decision y_i which is either $+1$ (not guilty) or -1 (guilty). Second, in a real, interactive group discussion, the individual group members reach a common decision y_g.

The real-world group decision y_g can be simulated by statistically aggregating the independently formed individual responses. MV simulates the group decision to be that of the majority of individuals, $y_g^{MV} = \text{sign}(\sum_{i=1}^{n} y_i)$. MV (as well as CWMV) assumes that individual responses are independent from each other given the ground truth, that is, individuals must form their decision only based on material that is not systematically shared between members. To illustrate a violation of this assumption, consider as another example a group of radiologists forming their individual diagnoses based on one and the same x-ray. They will not come to fully independent conclusions about the true state of the patient's condition because their opinions will be commonly influenced by the quality of the x-ray. In the worst case, multiple individual responses are fully dependent offering no more information than one single response. In our experiment, independence will be ensured by design in order to study CWMV—even though many real world situations will not allow for such a controlled environment.

When individuals report confidence ratings, c_i, MV can be improved upon

by using CWMV instead. These confidence ratings are assumed to be in the form of estimates for the probability of their decision being correct, $c_i = P(y_i \text{ is correct})$. In some situations, individuals can make such estimates (Griffin & Tversky, 1992; Martins, 2006; Regenwetter et al., 2014; Koriat, 2012a) and, under specific circumstances, assessing confidences is essentially the same as estimating the relative frequency of being correct (Brenner et al., 1996; Pouget, Drugowitsch, & Kepecs, 2016). CWMV transforms these confidences into optimal weights, which are the logarithmic odds (log odds), $w_i = \log(c_i/(1-c_i))$. See Nitzan and Paroush (1982) as well as Shapley and Grofman (1984), and find an intuitive account for using logarithmic odds as weights at the end of this section. Using these weights, CWMV simulates the group decision by

$$y_g^{\text{CWMV}} = \text{sign}\left(\sum_{i=1}^{n} w_i y_i\right) . \tag{3.1}$$

Similar to the individual confidence ratings, real groups can also report how confident they are in their group decision c_g. CWMV can simulate these group confidences based on the individual confidences by

$$c_g^{\text{CWMV}} = \frac{1}{1+\exp(-|\sum_{i=1}^{n} w_i y_i|)} . \tag{3.2}$$

To illustrate the computation of CWMV, consider again the three students deciding whether a suspect is guilty. Say, Student 1 votes for the suspect being innocent, $y_1 = +1$, but Students 2 and 3 believe the suspect to be guilty, $y_2 = -1$ and $y_3 = -1$. Aggregating these decisions using MV determines the simulated group decision to be guilty, $y_g^{MV} = \text{sign}((+1)+(-1)+(-1)) = -1$. Additionally, Student 1 reports being quite confident in their vote such that the probability of their judgment being correct is 76%, $c_1 = .76$. In contrast, Students 2 and 3 are very unsure with a confidence of only 51%, $c_2 = c_3 = .51$. Using CWMV to integrate these individual responses into a simulated group decision, the individual confidences are first transformed into weights with Student 1 having a higher confidence and, thus, a larger weight: $w_1 = \log(.76/.24) = 1.15$ versus $w_2 = w_3 = \log(.51/.49) = 0.04$. Then, CWMV leads to a different simulated group decision than MV finding the suspect not guilty, $y_g^{\text{CWMV}} = \text{sign}\left((+1.15)+(-0.04)+(-0.04)\right) = \text{sign}(+1.07) = +1$. Moreover, CWMV simulates the group's confidence in their verdict to be 75%, $c_g^{\text{CWMV}} = 1/[1+\exp(-|(+1.07)|)] = .75$. That is, the confident response from Student 1 is only slightly attenuated by the unconfident, opposing responses from Students 2 and 3 as might be realistic in a real group discussion. This example corresponds numerically to Scenario II from our experiment, which

we use to study in how far real groups are better represented by MV or CWMV simulations, see Table 3.1.

A technical note: The weights in CWMV are log odds because the logarithm is used as a convenient trick to transform a multiplication into a weighted sum. When computing the probabilities of a suspect being guilty or not, basic Probability Theory gives that odds ($o_i = c_i/(1-c_i)$) can be multiplied. In our example, the odds of the suspect being innocent are $o_1 = .76/.24$ and $o_2 = o_3 = .51/.49$. Multiplying these odds results in group odds, $o_g = o_1 \times o_2^{-1} \times o_3^{-1} = 3$ (o_2 and o_3 are inverted because Student 2 and 3 vote for guilty). Observe, that the group odds are indeed equivalent to the 75% group confidence computed by CWMV from above, $c_g/(1-c_g) = .75/.25 = 3$. By applying the laws of logarithm, multiplication of these odds is transformed into a sum of the log odds: $\log(o_1 \times o_2^{-1} \times o_3^{-1}) = (+\log(o_1)) + (-\log(o_2)) + (-\log(o_3))$, which allows to derive Equations 3.1 and 3.2. Note further that, when an individual is absolutely certain in their decision ($c_i = 0$ or $c_i = 1$), the odds o_i and weights w_i are undefined. In this case, by convention, the simulated group is set to be absolutely certain as well ($c_g = 0$ or $c_g = 1$). But if two participants came to opposite decisions and were both absolutely certain, by convention, their two responses would be discarded and the third individual's vote would decide (this situation did not occur in our experiment).

Given this formal framework of CWMV, the purpose of this study is to investigate how well individual responses (y_i and c_i) aggregated into simulated group responses (y_g^{CWMV} and c_g^{CWMV}) represent the real group responses from actual group discussions (y_g and c_g) on a trial-by-trial basis. We will modify Equations 3.1 and 3.2 using formal cognitive modeling in order to characterize how real groups deviate from these CWMV simulations.

3.3 Methods

Participants

A total of 21 participants (11 female, mean age = 21.4, range = 19 - 26) completed the experiment in seven groups of three. All were students who received either course credit for 30 min of participation or payment (4 EUR, equivalent to 4.5 USD). All participants had normal or corrected-to-normal vision and provided written informed consent prior to participation.

Stimuli & procedure

We adopted a procedure that has been established by Griffin and Tversky (1992) and extended it to a group setting. The experiment consisted of three practice trials followed by 12 experimental trials. Each trial consisted of an individual phase and a group phase, see Figure 3.1. Participants viewed rapid stimulus sequences consisting of 11 to 13 red and blue disks. Their task was to guess whether the stimulus sequence was generated by a fair coin (producing in expectation 50% red and 50% blue disks) or a biased coin (producing 60% red and 40% blue disks). Participants were instructed that both, the fair and the biased coin, are a priori equally likely. Griffin and Tversky (1992) showed that, in this task, participants' individual confidence ratings are well calibrated.

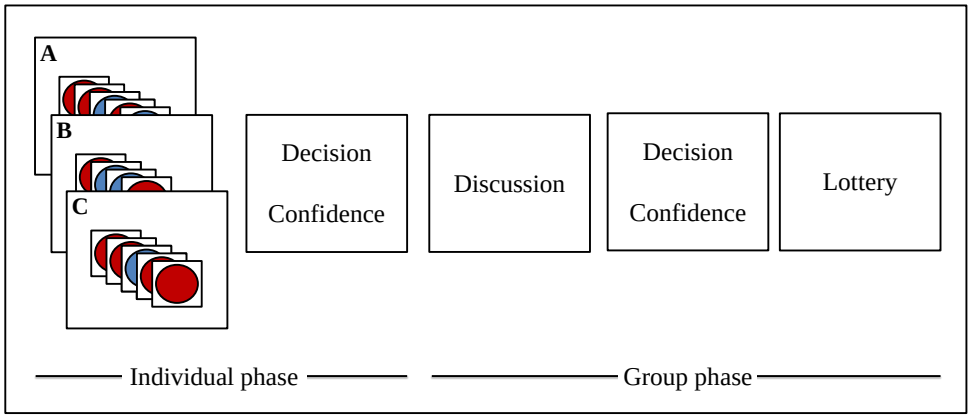

Figure 3.1: **Individual and group phase in each trial.** In the individual phase, each participant viewed different stimulus sequences consisting of 11 to 13 disks. Based on these sequences, individuals decided whether their sequence has more likely been produced by a fair coin (50% red, 50% blue) or a biased coin (60% red, 40% blue). Based on the ambiguity of the sequence, individuals reported a confidence in their own decision. In the group phase, participants combined their evidence into one group decision and confidence. In each trial, participants were incentivized for accurately judging their real group confidence using the Matching Probability method by Massoni et al. (2014).

Participants viewed different stimulus sequences simultaneously at individual laptops. Their viewing distance to the screen was approximately 60 cm. Each disk was presented for 100 ms with a diameter of 2.2 cm corresponding to a viewing angle of 2.1°. Disks were intermitted by a 100 ms blank interval creating the impression of a rapid stream. This presentation prevented partic-

ipants from performing explicit mathematical calculations so that they could only obtain an intuitive sense of confidence.

Depending on which coin better matched the stimulus sequence, participants made a decision for either the fair or the biased coin. Some stimulus sequences were more ambiguous than others providing different levels of confidence that participants reported on a visual analog scale from 50% ("I am completely unsure. The other option is equally likely.") to 100% ("I am completely sure. My decision is definitively correct."). Participants were instructed to report their subjective probability with which they believed their decision to be correct. In contrast to verbal scales, where participants give responses such as "somewhat likely" or "almost certain", this numeric scale is necessary because numeric values (here, c_i) are required to simulate group decisions (see Equation 3.1). However, it is noteworthy that a numerical scale can prompt participants to engage in formal thinking, which they would otherwise have not (Windschitl & Wells, 1996).

Table 3.1: **Ideal decisions and confidences.** In each trial, we applied one out of four scenarios (I–IV) which is defined by three stimulus sequences (A, B and C). Each of the three participants from a group viewed one stimulus sequence. Each individual stimulus sequence entails an ideal decision y_i^* and ideal confidence c_i^* that can be derived from probability computations. The ideal individual responses from each scenario determine the groups' ideal decision y_g^* and confidence c_g^* (see Methods section for an example calculation corresponding to Scenario II).

Scenario	Individual						Group	
	A		B		C			
	y_1^*	c_1^*	y_2^*	c_2^*	y_3^*	c_3^*	y_g^*	c_g^*
I	fair	87%	fair	70%	fair	62%	fair	96%
II	biased	76%	fair	51%	fair	51%	biased	75%
III	biased	88%	biased	54%	fair	81%	biased	66%
IV	fair	81%	biased	58%	biased	72%	fair	54%

The presented stimulus sequences determined ideal individual responses, which reflect posterior probabilities that can be computed using Probability Theory. Table 3.1 shows which responses the stimulus sequences would produce if participants were ideal observers. For example, assume that a participant saw the disk sequence red, red, blue, red and red. A fair coin would have

produced such a sequence with a likelihood of $p_{\text{fair}} = 0.5^5 = 3\%$ and the biased coin with $p_{\text{biased}} = 0.6^4 \cdot 0.4^1 = 5\%$. Because the biased coin was more likely to produce this stimulus sequence, the ideal decision is for the biased coin denoted by $y_i^* = +1$ (the asterisk denotes ideal values). The ideal confidence was $c_i^* = p_{\text{biased}}/(p_{\text{fair}} + p_{\text{biased}}) = 5\%/(3\% + 5\%) = 62\%$.

Note that scheduling individual reports before a group discussion (as in our experiment) improves group performance and prevents contamination of individual reports by the group decision (Sniezek & Henry, 1990). That is, individual reports remain independent because participants interacted only after they gave their individual responses.

After the individual phase, participants entered the group phase. Since participants had viewed different stimulus sequences that were produced by the same coin, they engaged in a group discussion to aggregate the individually gathered evidence and produce a real group response. Similar to the individual responses, groups reported a decision and rated their confidence in that decision. We label these responses based on real group discussions *reported group decision* and *reported group confidence* and later compare them to the *simulated group decision* and *simulated group confidence*, which we obtain from statistically aggregating individual responses using CWMV. Groups were allowed to give a group response not earlier than 30 seconds and discussions usually did not last longer than 2 minutes.

The ideal group responses, y_g^* and c_g^*, can be determined by adding the number of red and blue disks from all three stimulus sequences shown to the participants. Then, the same calculations as for ideal individual responses can be applied to compute the ideal group responses. Alternatively and equivalently, aggregating ideal individual responses using CWMV (Equations 3.1 and 3.2) also produces the ideal group responses because CWMV aggregates confidences in the theoretically correct way.

Across the 12 experimental trials, there were four Scenarios I – IV. Each scenario was defined by three stimulus sequences: A, B and C. Table 3.1 shows the ideal decision and confidence for each stimulus sequence in each scenario as well as the ideal group responses. Each participant saw one of those sequences from the current scenario. These scenarios were repeated three times in a randomized order for a total of 12 trials and the stimulus sequence that participants saw (A, B and C) were rotated so that participants viewed different stimulus sequences when a scenario was repeated. Importantly, Scenarios II and IV were designed so that MV and CWMV yield different predictions because the most confident individual should—according to CWMV—outweigh the relatively unconfident majority.

At the end of the group phase in each trial, the group was incentivized for

giving an accurate group confidence rating. They entered a lottery in which the group could win money depending on how accurate the reported group confidence was. This lottery, the Matching Probability method, was conceived by Massoni et al. (2014), see also Dienes and Seth (2010). The probability to win in this lottery is maximized if the group confidence is neither under- nor overestimated. Participants were instructed about the rules of this lottery and it was emphasized that chances to win are best when confidence ratings reflect the probability of the group decision to be correct. In each trial, groups could win 0.60 EUR (approximately 0.66 USD). Across 12 experimental trials, groups could win a total maximum of 7.20 EUR (7.90 USD) in addition to their compensation for participation. The sum was split equally among the three participants of the group. We did not apply this lottery for individual confidence ratings because these have already been shown to be reliable (Griffin & Tversky, 1992) so that incentivation was not necessary in the individual phase. In contrast, incentivation was applied in the group phase because we assumed that it is important to additionally motivate participants there and keep them engaged in the group discussions.

Formal cognitive modeling of CWMV

CWMV is the theoretically optimal way of aggregating individual responses. Real groups on the other hand may deviate from CWMV in various ways. To measure these deviations, we introduce four parameters into the CWMV framework in order to capture different aspects in which real groups deviate from CWMV:

- σ_i: *precision of individuals* in recovering the ideal confidence in their reported confidence ratings,
- β: *equality effect*, or, tendency of groups to weigh individual votes more equal or more extreme than CWMV would based on the individual confidences,
- γ: *group confidence effect* determining whether groups tend to over- or underestimate their confidences, and
- σ_g: *precision of groups* in determining the group confidence in accordance with CWMV simulations based on the individual confidence ratings.

We estimate individuals' precision, σ_i, in recovering the true strength of evidence of the displayed stimuli sequences. We assume that individuals are

not able to determine the ideal confidence but, instead, their actual responses will scatter around the ideal values. We describe this by an error term ϵ_i:

$$c_i = c_i^* + \epsilon_i.$$

This error term ϵ_i is normally distributed with mean zero (reflecting no absolute bias in individual confidence reports in accordance with Griffin & Tversky, 1992) and standard deviation σ_i. This standard deviation characterizes individuals' precision in recovering the true confidence. An ideal observer would be perfectly precise and make no errors, $\sigma_i = 0$, whereas larger values of σ_i indicate less precision.

Individuals make incorrect decisions if the actual confidence deviates below the 50% threshold resulting in the complementary confidence towards the incorrect decision.

$$y_i = \begin{cases} y_i^* & \text{(correct)}, \quad \text{if } c_i \geq 0.5. \\ -y_i^* & \text{(incorrect)}, \quad \text{otherwise.} \end{cases}$$

In our experiment, we used a half scale ranging from 50%–100% towards the decision made by the participant. For correct estimation, we transform the reported confidences into a full scale ranging from 0%–100% towards the correct decision (see Olsson, 2014) by inverting confidences towards the incorrect decision. For example, when an individual responded incorrectly with a confidence of 60%, we transform the confidence to $c_i = .4$ (40% towards the correct alternative) in order to estimate ϵ_i in each trial and thereupon σ_i.

Note that confidence ratings cannot be higher than 100%, which potentially causes a ceiling effect (Griffin & Brenner, 2004). However, in our experiment, ideal confidences for individual responses only range up to a maximum of 88% (Scenario III, Individual A in Table 3.1) so that there is enough room for positive deviations, ϵ_i, to avoid a large ceiling effect here.

Furthermore, we introduce the parameter β to estimate the equality effect capturing whether real groups weighted individual responses in a way that deviates from CWMV. This parameter acts upon the weights w_i as an exponent:

$$y_g^{\text{CWMV}}(\beta) = \text{sign}\left(\sum_i w_i^\beta y_i\right). \tag{3.3}$$

As the name suggests, the equality effect models groups assigning more equalized weights than naive CWMV, which is conceptually similar to the approach by Mahmoodi et al. (2015) but our model is technically different because we incorporate it into the CWMV framework. Here, the equality effect can vary between zero and infinity, $\beta \in [0; \infty]$. In the edge case of $\beta = 0$, every

weight would be transformed equally to $w_i^0 = 1$ producing the special case of (unweighted) MV. On the other hand, $\beta = 1$ would leave weights unchanged, $w_i^1 = w_i$, and would produce undistorted CWMV. Values in between, $0 < \beta < 1$, would represent some compromise in which individual confidences are considered to some extent but groups tend to equalize those weights. On the other side of the spectrum, larger values of $\beta > 1$ would represent an exaggeration of differences between weights so that the most confident individual's vote has a disproportionately large impact. In such situations, the most confident individual would tend to decide the vote single-handedly, which is equivalent to the predictions from maximum-confidence slating (Koriat, 2012b).

We additionally estimate whether groups under- or overestimate their group confidence, which is captured in the group confidence effect γ:

$$c_g^{\mathrm{CWMV}}(\beta, \gamma) = \frac{1}{1 + \exp(-\gamma |\sum_i w_i^\beta y_i|)} \ . \tag{3.4}$$

The group confidence effect allows for a non-linear scaling of the group confidences. This parameter can also vary between zero and infinity, $\gamma \in [0; \infty]$, where $\gamma < 1$ represents groups underestimating their confidence relative to the ideal statistical aggregation of individual responses, whereas $\gamma > 1$ represents an overestimation of group confidences. The special case of $\gamma = 1$ recovers undistorted (naive) CWMV.

Note that the equality effect β modifies individual weights and can potentially change the simulated group decision. In contrast, the group confidence effect γ only modifies a group's final confidence (hence, it does not appear in Equation 3.3 were the simulated group decision is determined). These two parameters capture deviations from naive CWMV simulations in a descriptive manner. For cautionary accounts against normative interpretations, see Gigerenzer (2018); Le Mens and Denrell (2011); and Neth, Sims, and Gray (2016).

Finally, we introduce an error term to the group confidence, $c_g = c_g^{\mathrm{CWMV}}(\beta, \gamma) + \epsilon_g$. This error term ϵ_g acts similar to the error term of individual confidence ratings. It is normally distributed with mean zero and standard deviation σ_g, where smaller values indicate higher precision of the group discussion process matching the ideal aggregation.

For estimation, the individual precision σ_i was measured by computing the average of sample variances across individuals and taking the square root. For the group parameters, we performed a grid search in which we varied β and γ in $[0, 2]$ and σ_g in $[0, 0.3]$ (larger values produced worse fits) with step sizes of 0.01. For each group, we chose the parameter combination that produced the maximum likelihood for the observed data using Equations 3.3 and 3.4 to predict the real group responses.

We validated this approach by conducting multiple parameter recovery simulations as suggested by Wilson and Collins (2019): We simulated data based on our model for fixed values of σ_i, β, γ and σ_g and demonstrated that our estimation approach recovered the ground truth parameters, see open material for details.

3.4 Results

We compared the average and median performance of real versus simulated groups, see Figure 3.2 and see Additional File 1 for estimates of each group. Real groups reported the correct (ideal) decision in 76.2% (SEM = 3.4%) of the trials (Mdn = 75.0%, IQR = 75.0–83.3). CWMV adequately simulated the average performance of real groups with 76.2% (SEM = 2.8%, Mdn = 75.0%, IQR = 70.8–83.3). In contrast, simulating group decisions using unweighted MV produced a lower accuracy of 66.7% (SEM = 3.6%, Mdn = 66.7%, IQR = 62.5–75.0) compared to CWMV with a mean difference of $M = 9.5\%$ ($SEM = 3.4\%$), $t(6) = 2.83$, $p = .030$. Comparing MV to real groups yielded a trend towards the same difference, $M = 9.5\%$ ($SEM = 4.6\%$), $t(6) = 2.07$, $p = .084$. We conducted two-sided, exact binomial tests to confirm this pattern: MV simulations were less accurate than CWMV simulations ($p = .016$) and real group decisions ($p = .016$).

Real versus ideal responses

Individual confidence ratings were well aligned with the ideal confidences, see Figure 3.3a. The average correlation between reported versus ideal confidences across individuals was $\bar{r} = .73$, 95% CI $[0.64, 0.80]$ (we used Fisher's z-transformation for combining correlations into averages). This finding replicates Griffin and Tversky (1992) showing that individual participants are able to evaluate the ambiguity in the presented stimulus sequences and report their confidences in form of subjective probabilities. Estimating the precision of individuals, we observed that reported confidences scattered around ideal confidences with a standard deviation of $\sigma_i = 13.3\%$, $SD = 6.6$, 95% CI $[9.8, 16]$.

However, confidence reports showed systematic deviations. In hard (difficult) trials with low ideal confidences, individuals overestimated those confidences. This is reflected in regression lines on average being at $M = 55\%$, 95% CI $[50.2, 58.8]$, where they should be at 50%. Additionally, high confidences were underestimated. The average slope of regression lines was lower than the ideal value 1, $\bar{b} = 0.78$, 95% CI $[0.61, 0.95]$. A slope of 1 would have indicated that ideal and reported confidences increased equally, whereas, here,

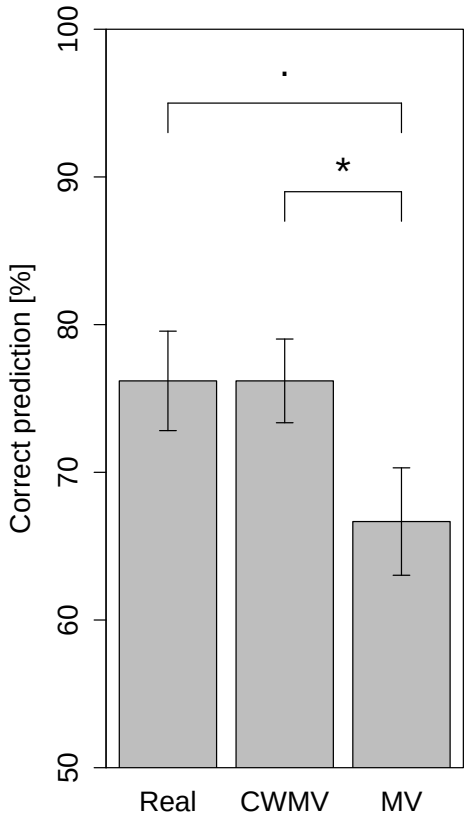

Figure 3.2: **Performance of real versus simulated groups.** Comparing the accuracy of real group decisions to simulated group decisions using either CWMV or MV for aggregation of the individual decisions and confidences. Simulated groups based on CWMV predict the performance of real groups very well, while simulated groups based on MV underestimate the performance of real groups. Error bars indicate standard errors of the mean computed across groups.
$^{*}p < .05$. $^{\cdot}p < .1$.

the observed slope below 1 indicated that increasing the true evidence strength from the presented stimulus sequences only led to a diminished increase in confidence.

Group confidence ratings showed a somewhat similar pattern, see Figure 3.3b (we again present median values). The average correlation between reported and ideal group confidences was high, $\bar{r} = .71$, 95% CI [.57, .80], Mdn = .71, IQR = .64–.79, but there was a relatively large root mean squared error, $RMSE = 0.16$. Real groups did not deviate from ideal values at low confidences: The regression lines at the ideal 50% were $M = 47\%$, 95% CI [40.7, 53.7],

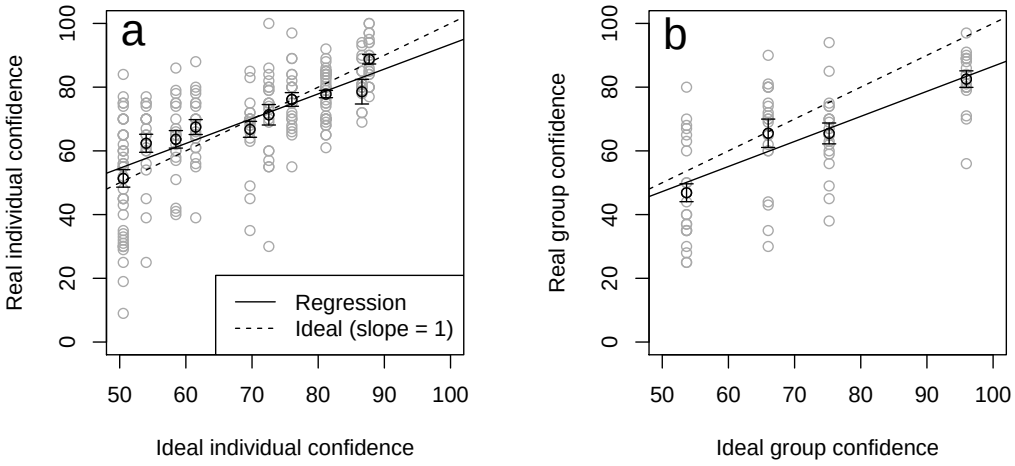

Figure 3.3: **Comparing ideal versus reported confidences from individuals and groups.** Ideal confidence (x-axis) range from 50% to 100% in accordance to Table 3.1. In contrast, reported confidences (y-axis) range from 0% to 100% because we flipped confidence ratings in cases where an incorrect decision was given (e.g., a reported confidence of 60% towards the incorrect decision is displayed as a confidence of 40% here). In (a), reported confidences from individuals (y-axis) are compared to the ideal values (x-axis; cf. c_1^*, c_2^*, and c_3^* from Table 3.1). Similarly in (b), reported confidences from groups (y-axis) are compared to the ideal values (x-axis; cf. c_g^* from Table 3.1). Black points indicate mean values—averaged across individuals in (a) and across groups in (b)—for each ideal value. Grey points indicate single trial responses. Error bars indicate standard errors of the mean.

Mdn = 48.6%, IQR = 42.5–50.6. Nevertheless, groups (similar to individual participants) underestimated high confidences resulting in an attenuated average slope relative to the ideal value of 1, $\bar{b} = 0.79$, 95% CI $[0.58, 0.99]$, Mdn = 0.77, IQR = 0.73–0.96. The large RMSE reflects this divergence for high confidences. Exact binomial tests confirmed these results: All groups had a correlation above 0 and a slope below 1, both $p = .016$, but intercepts scattered around 50%, $p = .336$. Note that we avoid common problems of regression in the context of over- versus underconfidence estimation since our regressions use the fixed ideal confidences as independent variables (x-axes in Figure 3.3), which exhibit no estimation error that would otherwise have lead to a biased analysis (Fiedler & Krueger, 2012; Olsson, 2014).

Real versus simulated group responses

Responses from real, interacting groups were well predicted by simulated responses using CWMV. Naive CWMV (Equations 3.1 and 3.2) produced an average correlation between reported and simulated confidences of $\bar{r} = .83$, 95% CI $[0.56, 0.94]$, Mdn = .82, IQR = .64–.92. Despite the high correlation, there was still a large discrepancy, $RMSE = 0.17$, reflecting deviations of real responses from naive CWMV, see Figure 3.4a.

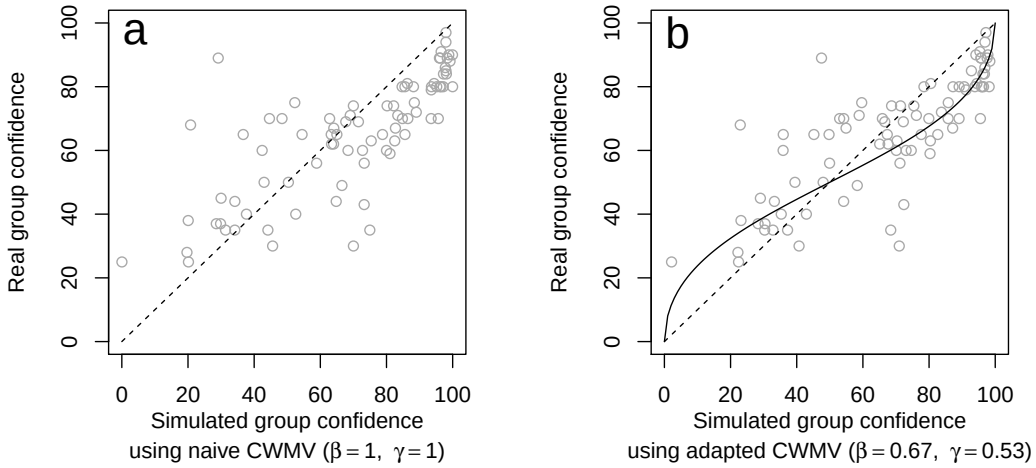

Figure 3.4: **Comparing real versus simulated group responses from statistically aggregating individual responses.** We used individual responses to simulate group confidences via CWMV (x-axis). These simulations predict responses from real, interacting groups (y-axis). In (a), we used naive CWMV as in Equations 3.1 and 3.2. The dashed line represents predictions from naive CWMV. This is equivalent to our formal cognitive modeling with equality effect $\beta = 1$ and group confidence effect $\gamma = 1$. In (b), we estimated the equality effect, $\beta = 0.67$, and group confidence effect, $\gamma = 0.53$, see Equations 3.3 and 3.4. This model predicts real group responses (solid line) but incorporates the fact that real groups treated individual votes more equal and displayed an underconfidence effect. In both subfigures, confidence ratings are inverted for incorrect responses. For example, the point (34%, 35%) in (a) corresponds to a trial with a simulated confidence of 66% and a reported confidence of 65% with both decisions being the same but incorrect, hence, both confidences were inverted.

We applied our formal cognitive model to estimate in how far real groups deviated from naive CWMV. The equality effect β was on average $M = 0.67$,

$SD = 0.30$, 95% CI [0.38, 0.95], Mdn = 0.74, IQR = 0.38–0.95. This indicates that groups used confidences similar to CWMV but tended towards equalizing those weights. Votes from confident individuals were given more impact on the group decision compared to unconfident individuals but not to the extent suggested by CWMV. We observed trials in which the most confident individual is overruled by the majority and the tipping point at which this happened was earlier than what naive CWMV simulations predict. This observation is captured in the equality effect estimate being smaller than 1, $\beta = 0.67 < 1$.

It is noteworthy that, since β estimates are always larger than zero, it is a priori expected to obtain an above zero average simply due to random errors. To account for this, we performed a randomization test where we randomly permutated individual confidences and estimated β from the resulting data set. Since the confidences in these randomized data sets are not indicative of the group's decision, the true equality effect is zero here. From 1,000 of such randomizations, we found that 95% of the obtained β estimates were below 0.4. This confirms that groups in our experiment (with $\beta = 0.67$) did take confidences into account ($\beta > 0$) but only to an attenuated extent ($\beta < 1$).

The group confidence effect γ was on average $M = 0.53$, $SD = 0.09$, 95% CI $[0.45, 0.61]$, Mdn = 0.62, IQR = 0.55–0.74, indicating that real groups tend to underestimate ($\gamma < 1$) their confidence compared to CWMV simulations based on the individual responses. In Figure 3.4b, this underestimation effect corresponds to a predicted curve (solid line) below the ideal values (dashed line).

The average group precision was $\sigma_g = 11\%$ (root mean square; $SD = 4$) with Mdn = 10%, IQR = 7%–12%. This precision of group confidences is comparable to the precision of individual confidences.

The adapted CWMV model using $\beta = 0.67$ and $\gamma = 0.53$ predicted confidences that are correlated with reported confidences to the same degree as naive CWMV, $\bar{r} = .84$, 95% CI $[0.68, 0.93]$, Mdn = .84, IQR = .72–.92. But in absolute terms, this adapted CWMV model matched the reported confidences better ($RMSE = 11\%$) than naive CWMV ($RMSE = 17\%$, mentioned above), $t(6) = 5.24$, $p < .002$, because the adapted model simulates group responses with an equality and underconfidence effect.

Note that going from Figure 3.4a to Figure 3.4b, points are shifted along the x-axis because the equality effect $\beta = 0.67$ changes simulated confidences and can even change the simulated decision (points crossing the 50% border in the x direction). The extent of these shifts depends on the exact constellation of individual confidences. On the other hand, the group confidence effect $\gamma = 0.53$ only maps the resulting, simulated confidences in a non-linear way to the reported confidences (solid curved line). This parameter reflects that

groups were less confident in their decisions than what naive CWMV predicted.

Model comparison of group response simulations

To evaluate our adapted CWMV model, we compared the full model to three special case models in which we fixed one parameter at a time (first $\gamma = 1$, second $\beta = 0$ and third $\beta = 1$). For this comparison, we computed the Bayesian Information Criterion (BIC; Schwarz et al., 1978). Using the Akaike Information Criterion (Akaike, 1973) instead of BIC yielded qualitatively identical results. Smaller BIC values indicate a better fit relative to the number of parameters in the model. For the full model, the total score (sum across groups) was $BIC_{\text{full}} = -101$.

As a first comparison, we pitch the full model against a model that fixes the group confidence effect $\gamma = 1$ but keeps the equality effect β free. This model assumes that groups may only deviate from naive CWMV in terms of how they assign weights to the individual votes but exhibit no general over- or underconfidence. Here, the total score was $BIC_{\gamma=1} = -59$ indicating a worse fit as compared to the full model. The Bayes factor resulting from the BIC scores of the two models (e.g., see Farrell & Lewandowsky, 2018, chapter 11) clearly supported the full model, $BF_{\text{full}/\gamma=1} > 1000$. This supports the notion that group responses are best characterized by an overall underconfidence effect.

The second comparison fixes $\beta = 0$ but keeps γ free. This model is equivalent to an (unweighted) MV with group confidence effect. Here, the total score was $BIC_{\beta=0} = -63$, again supporting the full model, $BF_{\text{full}/\beta=0} > 1000$. This indicates that participants incorporate confidence ratings in the group discussion.

For the third comparison, we fix $\beta = 1$: This model assumes that real groups weigh individual votes exactly according to undistorted CWMV but still allows for an overall confidence effect of the group since γ is free. This model was on par with the full model, $BIC_{\beta=1} = -102$, with an inconclusive Bayes factor, $BF_{\text{full}/\beta=1} = 0.71$. This indicates that, according to the BIC criterion, fixing $\beta = 1$ did not perform worse (when accounting for the additional free parameter) than the full model, which keeps β free. On the other hand, when performing a model fit comparison irrespective of the number of parameters (Farrell & Lewandowsky, 2018, chapter 10), the full model performs better than that with fixed $\beta = 1$, $\chi^2(7) = 16.9$, $p = .018$. To confirm that incorporating the equality effect β as a free parameter in our model conveys an advantage even when weighing parsimony against model fit, future research with increased sample sizes are necessary.

3.5 Discussion

Including confidence ratings in the theoretically optimal way using CWMV increases the simulated group performance over MV. Real groups are more accurately represented by CWMV when individuals provide reliable and independent confidence ratings. Even though real groups consider confidence ratings similar to CWMV, they tend to treat individual responses more equally giving more confident individuals less impact on the group decision than naive CWMV simulations, which is consistent with an equality bias (Bang & Frith, 2017; Mahmoodi et al., 2015). Additionally, groups tend to underestimate their confidences.

In our study, individuals were overconfident in hard (difficult) trials and underconfident in easy trials—a finding often referred to as the hard-easy effect (Gigerenzer et al., 1991). Hard trials allow participants to make a correct decision about 50% of the time but reported confidences were larger. In contrast, easy trials allow for close to 100% confidences but reported confidences were strictly lower. This hard-easy effect, or underextremity (Griffin & Brenner, 2004), can be explained by a regressive tendency (Moore & Healy, 2008). That is, participants were biased towards their prior belief to observe trials with moderate difficulty. But it can also be explained by a bias introduced through the response format: Olsson (2014) argue that a half scale (50%–100%), as it is often used, biases participants to respond closer to the center of the scale.

In contrast, real groups did not tend to be overconfident for hard trials in our setting but real groups exhibited overall underconfidence in a double sense: First, group confidences were lower than ideal responses (see Figure 3.3b). Second, group confidences were lower than determined by CWMV simulations based on individual responses (see Figure 3.4b).

Interestingly, confidence ratings reflected subjective probabilities rather than consistency in our study. For example, we presented a stimulus sequence that is suited to evoke a low ideal confidence of 54% (see Table 1, Scenario III, Individual B). For this sequence, participants gave the correct decision in 85.7% of the trials and reported confidences relatively close to the ideal confidence with an average of 62% (see Figure 3.4, second black point from left). In other words: Participants consistently determined the correct decision but nevertheless reported in their confidence ratings that the strength of evidence was quite low as intended.

One limitation that our well controlled setting cannot account for are situations in which individuals consensually reach incorrect decisions with high confidence (see Koriat, 2015, 2017; Litvinova et al., 2019). In such situations,

confidences towards the incorrect decision are aggregated and can lead to high group confidences towards incorrect decisions. In how far CWMV can adequately reflect real groups in these situations remains to be shown because consensually incorrect decisions were too rare in our setting to allow inferences, see bottom left quadrants in Figure 3.4.

Further insight into group processes can be gained by fixing the ideal group confidence and varying the constellation of individual confidences. For example, our Scenario II determined an ideal group confidence of 75% based on one confident individual (76% for biased coin) and two almost uninformative individuals (51% for fair coin). The same ideal confidence of 75% would come from three equally confident members (59% for biased coin). CWMV predicts the same ideal confidence but real groups may behave differently in these two cases. From our current estimates of the equality bias ($\beta = 0.67$), we predict that real groups are more confident in the latter constellation where each individual contributes an equal confidence as compared to a situation where only one individual is very confident.

Our controlled setting provided optimal conditions for CWMV with independent confidence ratings but it was rather artificial. This allowed us to verify that groups are indeed able to perform confidence weighting to some extent. However, in real world tasks, bad calibration of confidences may prevent simulated groups to perform as well as real groups. For example, Klein and Epley (2015) observed that individuals could not report well calibrated confidence ratings but real groups still outperformed simulations using MV. One possibility is that individuals failed to rate their confidence in a comparable way when verbal scales were used instead of numeric scales (Windschitl & Wells, 1996): Klein and Epley used a 9-point Likert scale from "not at all confident" (1) to "very confident" (9). Nevertheless, participants might have been able to share calibrated confidences in the real group discussions. This could have led to a better performance of real compared to simulated groups.

Another possible reason for real groups outperforming simulated groups is that the assumption of independence is violated. These—arguably more realistic—situations have been investigated under the name of *hidden profiles*, where a hidden profile determines the distribution of information that is either common among or unique to individuals (Stasser & Titus, 2003; Stasser & Abele, 2020). Distinguishing between evidence that is held by all individuals of a group versus evidence that is uniquely known by few individuals is a crucial aspect of successful real groups (Mercier, 2016). Consider again the example of three individuals deciding whether a suspect is guilty or not. Say, individuals have in total five pieces of evidence: two incriminating, I_1 and I_2, and three exonerating, E_1, E_2 and E_3. All individuals know all the incriminat-

ing evidence but each individual knows only one unique piece of exonerating evidence. That is, the first individual knows I_1, I_2, and E_1; the second knows I_1, I_2, and E_2; and the third knows I_1, I_2, and E_3. For each individual there is more incriminating evidence and each would decide 'guilty' with some confidence. Incorrectly assuming independence, CWMV would simulate the group decision to be guilty as well. However, a real group might lay out all the evidence, find in total more exonerating evidence, and decide 'not guilty'.

There are some approaches to handle such dependencies formally (Kaniovski & Zaigraev, 2011; Shapley & Grofman, 1984; Stasser & Titus, 1987) each coming with its own set of particular, additional assumptions. To sketch the approach that we find most promising: CWMV could be applied not to the potentially dependent individual responses but to the independent pieces of evidence, with confidences indicating the strength of each piece of evidence. Incorporating CWMV in this way could improve theoretical predictions: Rather than comparing group performance to the best individual (as is often done), CWMV-inspired approaches may provide a more adequate baseline for group performance even when information is distributed in a way that violates the independence assumptions for individual responses.

Conclusion

Confidence ratings of individuals play an important role in real group decisions and can be used to increase simulated group performance. In a controlled setting, real groups have proven to aggregate confidences in a way that is to some extent consistent with the CWMV even though they tend to treat individual responses more equal and with lower confidence than when using CWMV simulations. Developing group simulation methods (for example to account for dependencies) and comparing simulated group decisions using those methods to real group decisions will deepen our understanding of real world group discussions.

Chapter 4

Ensemble Performance Bounds

Manuscript published on preprint server: Meyen, S., Göppert, F., Alber, H., von Luxburg, U., & Franz, V. H. (2021) Specialists Outperform Generalists in Ensemble Classification. *arXiv preprint arXiv:2107.04381*. `https://arxiv.org/abs/2107.04381`

Abstract

Consider an ensemble of k individual classifiers whose accuracies are known. Upon receiving a test point, each of the classifiers outputs a predicted label and a confidence in its prediction for this particular test point. In this paper, we address the question of whether we can determine the accuracy of the ensemble. Surprisingly, even when classifiers are combined in the statistically optimal way in this setting, the accuracy of the resulting ensemble classifier cannot be computed from the accuracies of the individual classifiers—as would be the case in the standard setting of confidence weighted majority voting. We prove tight upper and lower bounds on the ensemble accuracy. We explicitly construct the individual classifiers that attain the upper and lower bounds: specialists and generalists. Our theoretical results have very practical consequences: (1) If we use ensemble methods and have the choice to construct our individual (independent) classifiers from scratch, then we should aim for specialist classifiers rather than generalists. (2) Our bounds can be used to determine how many classifiers are at least required to achieve a desired ensemble accuracy. Finally, we improve our bounds by considering the mutual information between the true label and the individual classifier's output.

4.1 Introduction

Suppose a black-box classifier returns a prediction along with a confidence value indicating the probability that this prediction is correct. For example, a deep neural network may take an image of a patient's retina and predict whether the patient suffers from a retinal disease (example taken from Ayhan et al., 2020; Leibig, Allken, Ayhan, Berens, & Wahl, 2017; Ayhan & Berens, 2018). It also outputs a confidence in this prediction based on the particular retina image. Suppose we apply k black-box classifiers each receiving its own retina image as input, observe their individual prediction-confidence output pairs and combine them into an ensemble classifier. In this paper, we investigate the question: **Given that the individual black-box classifiers output predicted labels together with confidences, what can we say about the accuracy achieved by the ensemble classifier? And can we characterize which type of individual classifier leads to a better vs. worse ensemble performance?**

At first glance, this problem seems trivial. If we know that all individual classifiers have the same accuracy, then the best we can do is a *majority vote* (MV; Grofman et al., 1983; De Condorcet et al., 2014). This is common practice in many ensemble approaches in machine learning, for example in random forests (Breiman, 2001). If some of the individual classifiers are known to have a higher accuracy than others, they should receive a higher weight. Based on this knowledge, the best we can do is *confidence weighted majority voting* (CWMV; see Nitzan & Paroush, 1982; Einhorn et al., 1977), where the confidence in a classifier is derived from its overall accuracy. In both cases, the accuracy of the ensemble classifier is well known and can be computed from the accuracies of the individual classifiers (under mild assumptions such as conditional independence).

However, these approaches do not fully capture the retina example from the beginning because they do not consider the classifier's "local confidences": For each image, the classifier produces a confidence in its prediction for this particular image. And this is where it gets interesting: Instead of using CWMV, where the confidence is based on the overall accuracy of the classifier, we get better classification results by using the local confidences for each prediction. Somewhat surprisingly, in this setting we can no longer compute exactly what the resulting ensemble accuracy is going to be. On the contrary. We prove in Section 4.4 that, depending on the distribution of confidence values, there is a whole range of ensemble accuracies that can occur. **Our contribution** is to derive lower and upper bounds on the ensemble accuracy in this setting. This is interesting if we want to determine how many classifiers (each requiring an

independently drawn, potentially costly retina image) are needed to guarantee a certain ensemble accuracy. From our proofs, we derive guiding principles on how to construct ideal classifiers for an ensemble: We will show that it is better to include "specialist" classifiers that are particularly good on some instances and close to random guessing on others rather than to include "generalist" classifiers that are moderately good on all instances. This is true for independent specialists that did not coordinate to specialize on distinct subsets of the input space.

In the second part (Section 4.5), we additionally look into the mutual information as an indicator for the effectiveness of a classifier in ensembles. We provide better bounds on the possible ensemble accuracies. Even when classifiers have the same accuracy, they can differ in how much information they provide about the true label and therefore differ in their contribution to the ensemble.

Of course, ensemble methods are abundant in machine learning and statistics, just consider random forests, bagging and boosting as examples. Compared to these lines of work, our approach starts from the other end. Rather than explicitly training certain ensembles, we are looking for generic building principles for ensembles. We build on the setting of probability elicitation DeGroot & Fienberg, 1983; Masnadi-Shirazi, 2013 and ask the question: Which possible ensemble accuracies can be achieved by a set of individual classifiers with known accuracies, and which kind of individual classifiers produce the best- and worst-case ensembles?

4.2 Setup, Notation, and Background

4.2.1 Individual Classifiers and Confidences

We work in a standard classification setting with input points X in some abstract input space $\mathcal{X}$, binary labels $Y \in \{-1, 1\}$, and a joint probability distribution P on $\mathcal{X} \times \mathcal{Y}$. We assume that both labels have the same probability, $P(Y = +1) = P(Y = -1) = 0.5$, meaning that, in our retina example, patients equally often have the disease as they do not have the disease (we make this assumption to keep the notation simple but our results may be generalized to a setting with unequal probabilities). A black-box classifier, upon observing a test point $X \in \mathcal{X}$ (a retina image), outputs two quantities: the predicted label $\hat{Y} \in \{-1, 1\}$ and the pointwise, or local, confidence $C \in [0.5, 1]$. We assume Bayes classifiers so that the predictions are optimal, $\hat{Y} = \operatorname{argmax}_{y \in \{-1,+1\}} P(Y|X)$. Furthermore, we assume the classifiers to be perfectly calibrated. That is, the local confidence is exactly the probability of

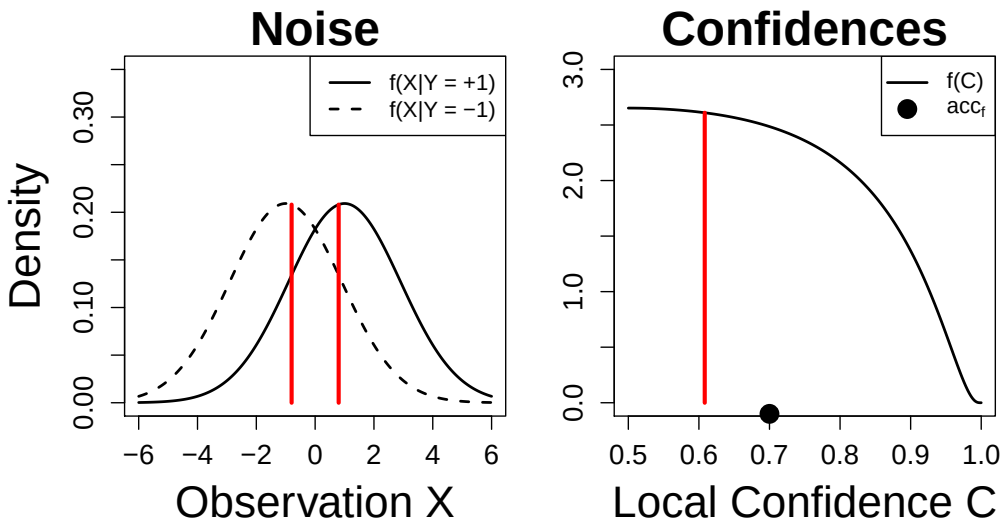

Figure 4.1: **Example of a confidence distribution.** A classification setting with normal noise distribution (left) is mapped to a confidence distribution (right). For normally distributed noise with $\sigma = 2.1$, an observation of $X = 0.8$ (indicated by the red bar) corresponds to a local confidence of $C(X) = 61\%$ such that an individual classifier outputs $(\hat{Y} = +1, C = 61\%)$. Therefore, the density at $X = 0.8$ (plus that at $X = -0.8$, because it produces the same confidence albeit predicting $\hat{Y} = -1$) corresponds to the density at confidence $C = 61\%$ on the right. The overall accuracy of the black-box classifier in this setting is $\text{acc}_f = 70\%$, indicated by the black dot.

that particular prediction to be correct, $C(X) = P(Y = \hat{Y}|X)$, or short, C. (The range of possible ensemble accuracies would be even larger if we dropped this calibration assumption.) We call C a *local* confidence to stress that it is different for each input point X whereas we use the term *accuracy* to refer to the overall probability of a classifier making a correct prediction across the input space.

In the following, we will describe a black-box classifier by its **confidence distribution** $f(C)$. A confidence distribution f is a probability distribution on $[0.5, 1]$ that describes how often each local confidence C is sampled. In Figure 4.1, we show how a confidence distribution is related to a classification setting with normal noise. In our retina example, a classifier's confidence distribution describes how often we get retina images of a certain quality such that a classifier can make a prediction with confidence C. We assume the confidences to be independent of the true label, $f(C|Y) = f(C)$ meaning that the quality of a retina image is independent of whether the patient has the disease or not.

If the confidence distribution f of a classifier is known, its classification accuracy can be computed from f:

$$\text{acc}_f := P(Y = \hat{Y}) = \int_{0.5}^{1} f(c) \cdot c \; dc \; . \tag{4.1}$$

However, in this paper, we deal with the more realistic scenario where the underlying local confidence distribution f is unknown and we only know the accuracy acc_f of an individual classifier.

4.2.2 Ensemble Classifiers

We obtain an ensemble prediction by optimally combining the outputs of k individual classifiers. The individual classifiers have unknown confidence distributions f_1, f_2, ..., f_k. We will make the important assumption that these confidence distributions are pairwise independent, $\forall i \neq j : f_i \perp f_j$. In our example, this means that the quality of retina images is independently drawn for each classifier (from its unknown confidence distributions). Under this assumption, the individual confidence distributions combine into the ensemble confidence distribution, denoted by f_e (see Section 4.3.2). Since we do not know the individual confidence distributions, f_1, f_2, ... f_k, we also do not know the exact ensemble confidence distribution, f_e.

The **goal of this paper** is to determine the accuracy of that ensemble classifier, acc_{f_e}, given that we only know the accuracies of the individual classifiers acc_{f_1}, acc_{f_2}, ... acc_{f_k} but not their exact confidence distributions, and to characterize which type of individual classifier leads to a better / worse ensemble performance.

4.3 Confidence Weighted Majority Voting

In this section, we first recap the traditional approach of CWMV and then introduce our modification based on local confidences, which we call *l*CWMV.

4.3.1 Traditional Approach: CWMV

In the traditional setting of CWMV (Grofman et al., 1983; Nitzan & Paroush, 1982), upon receiving input, a classifier outputs a prediction $\hat{Y}$, but not the local confidence for the particular test point. All we know is the (global) accuracy of the black-box classifier. In an ensemble, we observe a set of k predictions $\hat{Y}_1$, $\hat{Y}_2$, ..., $\hat{Y}_k$ from classifiers with accuracies acc_1, acc_2, ..., acc_k. It has been proven (Grofman et al., 1983) that the optimal way to form an ensemble

prediction in this scenario is to weight the individual classifiers' votes based on their accuracies, $W_i = \log(\text{acc}_i/(1-\text{acc}_i))$. These weights are therefore based on the overall accuracies of the individual classifiers. Traditional CWMV then produces the optimal ensemble prediction $\hat{Y}_e$ and the ensemble confidence in that prediction C_e as

$$\hat{Y}_e = \text{sign}\left(\sum_{i=1}^{k} W_i \hat{Y}_i\right), \text{ and} \tag{4.2}$$

$$C_e = \left(1 + \exp\left(-\left|\sum_{i=1}^{k} W_i \hat{Y}_i\right|\right)\right)^{-1}. \tag{4.3}$$

Note that when $\text{acc}_i = 1$ for any i, the weight W_i is undefined and therefore $\hat{Y}_e$ and C_e are set to $\hat{Y}_e = \hat{Y}_i$ and $C_e = 1$ by convention because classifier i is always correct in its prediction.

4.3.2 Modification With Local Confidences: *l*CWMV

We modify the traditional setting such that, upon receiving input point X, a classifier outputs its prediction $\hat{Y}$ together with a local confidence $C(X)$. It is straightforward to see that the optimal combination of the outputs of k classifiers, $i \in \{1..k\} : \left(\hat{Y}_i, C(X_i)\right)$, will base the weights not on the accuracies of the individual classifiers but on their local confidences for their individual input points: $W(X_i) = \log(C(X_i)/(1-C(X_i)))$. The ensemble prediction $\hat{Y}_e$ and confidence C_e are then computed analogously to Equations (4.2) and (4.3), using the local weights $W_i = W(X_i)$. While the weights were constant in traditional CWMV, they can differ from prediction to prediction in *l*CWMV.

In contrast to the traditional approach, we can no longer compute the ensemble accuracy, acc_{f_e}, based on the accuracies, acc_1, acc_2, ..., acc_k. Only when the exact distributions over the local confidences, f_1, f_2, ..., f_k, are known, we can derive the confidence distribution of the ensemble, f_e, and thereupon the ensemble accuracy acc_{f_e}. In the following, we denote the operation of combining individual confidence distributions in the *l*CWMV setting by $\otimes$ (formally defined in the proofs, Section 4.6) so that the ensemble confidence distribution is denoted by $f_e := \otimes_{i=1}^{k} f_i$.

In many practical examples, the confidence distribution of individual classifiers will not be known. Especially in cases where there is a high cost for obtaining predictions, as in our retina example, estimating confidence distributions is expensive. Not knowing the individual classifier's confidence distribution but only their overall accuracies entails some uncertainty about the ensemble confidence distribution f_e. Consequently, there are different possible

values for the ensemble accuracy, acc_{f_e}. The question we now answer is: What are the best and worst ensemble accuracies that can be achieved? And which individual confidence distributions contribute more to the ensemble accuracy than others?

4.4 Individual accuracies do not uniquely determine ensemble accuracy

In this section, we provide bounds on the ensemble accuracy when only the accuracies of the individual classifiers are known. The relevant aspect of a classifier will be its confidence distribution, $f(C)$, which is only constraint by the given individual accuracy. The classifiers that produce the best- and worst-case ensemble accuracies will be called specialists and generalists. A specialist and a generalist, even when they have the same accuracy, behave very differently in ensembles due to their different confidence distributions, see Figure 4.2 (top). Because these two extreme classifiers produce only a discrete amount of confidence levels, we will denote their probability distributions as weighted sums of Dirac probability measures δ_c that have point mass 1 at point c.

Intuitively, a classifier is a specialist if it achieves high confidence on some parts of the input space while it is close to random guessing on the rest. Formally, it outputs predictions with confidence either $C = 50\%$ (random guessing) or $C = 100\%$ (absolute certainty), see Figure 4.2 (top left). The proportion of these two cases determines the overall accuracy of the specialist.

Definition 1. ***(Specialist)*** *A binary black-box classifier with accuracy acc is called specialist if its confidence distribution is given by*

$$f_{acc}^{specialist} = w_{0.5}\delta_{0.5} + w_1\delta_1,$$

with constants $w_{0.5} = 2(1 - acc)$ *and* $w_1 = 2(acc - 0.5)$.

Generalists, on the other hand, work equally well on all of the input space. Their confidence is constant with $C = \text{acc}$, see Figure 4.2 (top right).

Definition 2. ***(Generalist)*** *A binary black-box classifier with accuracy acc is called generalist if its confidence distribution is given by*

$$f_{acc}^{generalist} = \delta_{acc}.$$

The following theorem states that generalists and specialists are the worst and best case classifiers when used in an ensemble.

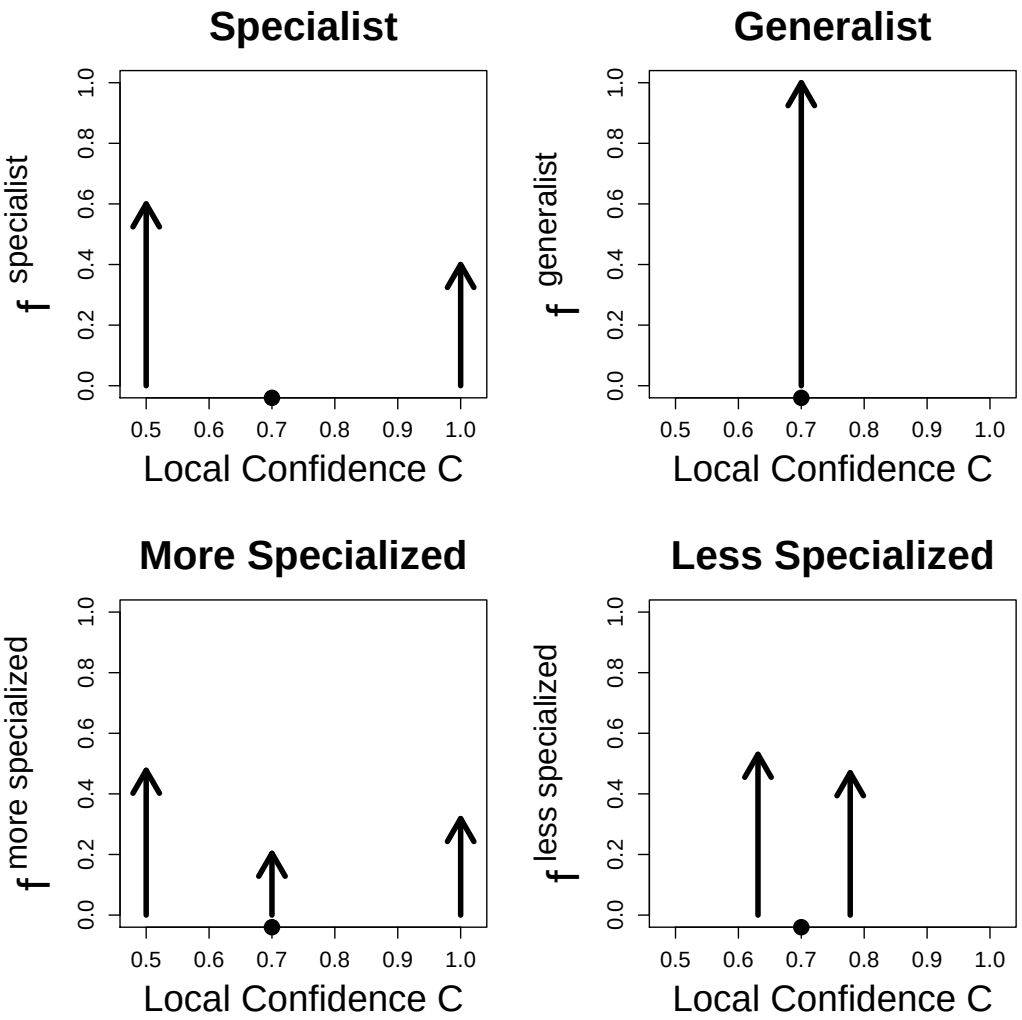

Figure 4.2: **Classifiers' confidence distributions that provide best- and worst-case ensemble accuracies.** Top row: Given an individual classifier with accuracy $\text{acc}_f = 70\%$ (black dot), the plots show the confidence distribution of a corresponding specialist and generalist classifier. The confidence distributions consist of point masses as indicated by the arrows (cf. Definitions 1 and 2, Section 4.4). Bottom row: Confidence distributions of more and less specialized classifiers to same accuracy of $\text{acc}_f = 70\%$ and the information $\text{I}_f = 0.25$ bit (cf. Definitions 5 and 6, Section 4.5.2).

Theorem 3. ***(Specialists and generalists bound the ensemble accuracy)*** *Consider k classifiers with individual accuracies acc_i and confidence distributions f_i ($i \in \{1..k\}$). For each i, let $f_{acc_i}^{generalist}$ and $f_{acc_i}^{specialist}$ be a gener-*

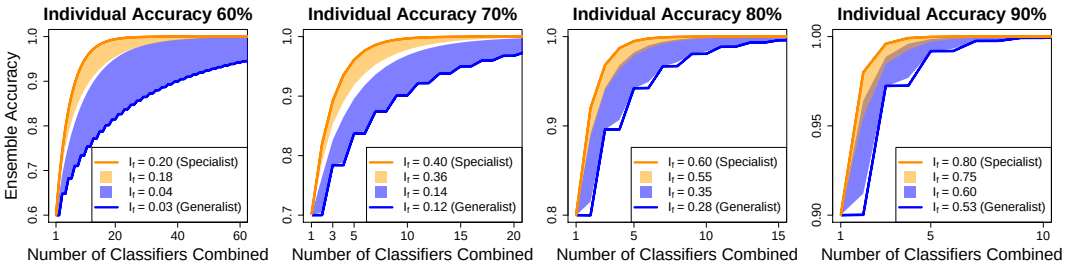

Figure 4.3: **Illustrations of the bounds in Theorems 3 and 7**. Each plot shows the ensemble accuracy acc_{f_e} as a function of the number k of individual classifiers. Within each subplot, all individual classifiers have the same individual accuracy acc_i as indicated in the title of the subplot. Best- and worst-case ensemble accuracies according to Theorem 3 are shown as solid orange and blue lines, achieved by the two extreme cases, specialists and generalists, which also have the highest resp. lowest information (Theorem 4). When in addition to the accuracy the information I_f of the individual classifiers is known, the range of possible ensemble accuracies gets smaller (Theorem 7): For example, individual classifiers with a relatively high information produce a range of possible ensemble accuracies (light orange shaded areas) exceeding that of classifiers with low information (light blue shaded areas). The steps in the lower bound result from the problem of conducting majority votes in ensembles of even size (no tiebreakers).

alist resp. specialist classifier that has the same accuracy as classifier i. Now consider the ensemble classifier based on the original classifiers with ensemble confidence distribution $f_e = \bigotimes_{i=1}^{k} f_i$ according to lCWMV as well as the ensemble of generalists and ensemble of specialists with ensemble confidence distributions $f_e^{generalist} = \bigotimes_{i=1}^{k} f_{acc_i}^{generalist}$ and $f_e^{specialist} = \bigotimes_{i=1}^{k} f_{acc_i}^{specialist}$. Then the accuracy of the original ensemble is lower and upper bounded by the accuracies of the generalist and specialist ensembles:

$$acc_{f_e^{generalist}} \leq acc_{f_e} \leq acc_{f_e^{specialist}} \ .$$

The formal proof of this theorem is in Section 4.6.3. The proof idea is that merging confidence distributions makes the ensemble accuracy worse: When a classifier does not distinguish between high vs. low confidence cases (Figure 4.2 top left) and instead always outputs an average confidence (Figure 4.2 top right), the ensemble is less effective in weighing that classifier's predictions. It helps to know which predictions should be taken into account (high confidence cases) and which should be disregarded (low confidence cases). Distinguishing between high and low confidence cases is related to the concept of refinement

(DeGroot & Fienberg, 1983; Masnadi-Shirazi, 2013), see Section 4.6.2. In consequence, the best ensemble accuracy comes from the most refined confidence distributions (specialists); and the worst ensemble accuracy comes from the least refined confidence distributions (generalists). Even though confidences can vary from prediction to prediction in our *l*CWMV setting, generalists do not make use of this possibility and always output the same confidence. They receive a constant weight as in the traditional CWMV setting. Therefore, our lower bound for the ensemble accuracy corresponds to the behavior of traditional CWMV.

To get an intuition for the meaning of the theorem, consider again the retina example. Assume we have $k = 3$ classifiers that take independently drawn retina images from a patient and return predictions as well as confidences. Let their predictions be correct with accuracies $\text{acc}_1 = \text{acc}_2 = \text{acc}_3 = 70\%$ (in Figure 4.3, second plot). Then, if these classifiers are generalists, their ensemble accuracy will be 78% (blue lower bound). But if they are specialists, their ensemble accuracy will be 89% (orange upper bound)—a large range that makes a crucial difference in practice. If the three classifiers' confidence distributions are not that of specialists or generalists (as in Figure 4.2) but an intermediate case as in our normal noise example (Figure 4.1, right) the ensemble accuracy is in between the bounds, here, at 82%. See Figure 4.3 for more numerical examples. We only show cases in which the individual accuracies are equal but our theorems can be applied to classifiers with different individual accuracies.

Theorem 3 carries two important messages: (1) Even when we know the accuracies of the individual classifiers and we combine their output in the statistically optimal way (with *l*CWMV), we are far from being able to predict the ensemble accuracy (unless we know the confidence distributions). (2) When we use ensemble methods and have the choice to construct our individual classifiers from scratch, then we should aim for specialist classifiers rather than generalists.

Crucially, **even without coordination between the classifiers, specialization is advantageous.** Specialists' confidence distributions are, by assumption, independent. Specialists do not divide the input space by specializing on separate regions. In our retina example, it is *not* the case that one specialist is trained on one subtype of retinal disease while a different specialist is trained on another subtype. This would contradict our assumption that individual confidences are independently drawn ($\forall i, j \in \{1..k\} : f_i \perp f_j$, introduced in Section 4.2.2). When one specialist classifier achieves a high confidence it is *not* more likely that the other specialists produce a low confidence as it would be the case when they had separate specializations. This

highlights the effectiveness of specialists even in independent ensembles.

4.5 Better Bounds for Ensemble Accuracy With Mutual Information

As shown, the range of possible accuracies of ensemble classifiers outlined in Theorem 3 can be large. In this section, we improve the bounds to better predict what the ensemble accuracy will be. We will assume that another performance measure next to the individual classifier's accuracy is known: the mutual information between the true label and the individual classifier's output (Shannon, 1948; Cover & Thomas, 2006; MacKay, 2003). This is just one alternative quality measure of the classifier, and many more such scoring functions exist (see Masnadi-Shirazi, 2013, 2017). We choose the mutual information for its natural properties but our results can be transferred to other convex scoring functions.

4.5.1 Mutual Information Measures Effectiveness in Ensembles

In addition to the accuracy of a classifier, we consider the mutual information I between the true label Y and the classifier's output $O = (\hat{Y}, C)$, which is $I(Y;O) = H(Y) - H(Y|O)$, where H denotes the (conditional) entropy of a random variable. With some simple rearrangement (see Section 4.6.1), the mutual information can be shown to only depend on the classifiers' confidence distribution f:

$$\mathrm{I}_f := I\left(Y;O\right) = \int_{0.5}^{1} f(c) \cdot \left(1 - H_2(c)\right)\ dc, \tag{4.4}$$

where H_2 is the binary entropy, $H_2(c) = c \log_2\left(\frac{1}{c}\right) + (1-c)\log_2\left(\frac{1}{1-c}\right)$ for $c \in [0.5, 1]$. In the following, we will denote a classifier's *information* by I_f, analogously to its accuracy acc_f, as a performance measure based on a classifier's confidence distribution f. For classifiers with fixed accuracy acc_f, specialists have the highest possible information and generalists have the lowest possible information.

Proposition 4. ***(Specialists and generalists bounds the individual information)*** *A classifier with confidence distribution f and accuracy acc has an information between*

$$I_{f_{acc}^{generalist}} \leq I_f \leq I_{f_{acc}^{specialist}}.$$

The proof is in Section 4.6.3. In our example, when an individual classifier has an accuracy of acc = 70%, its transmitted information lies between 0.12–0.4 bit, depending on its confidence distribution. With this, all classifiers can be described by two values, their accuracy acc and information I, and these values lie in the shaded area in Figure 4.4 (middle): Higher accuracy (along the x-axis) loosely coincides with higher information (y-axis) but this is no one-to-one relation.

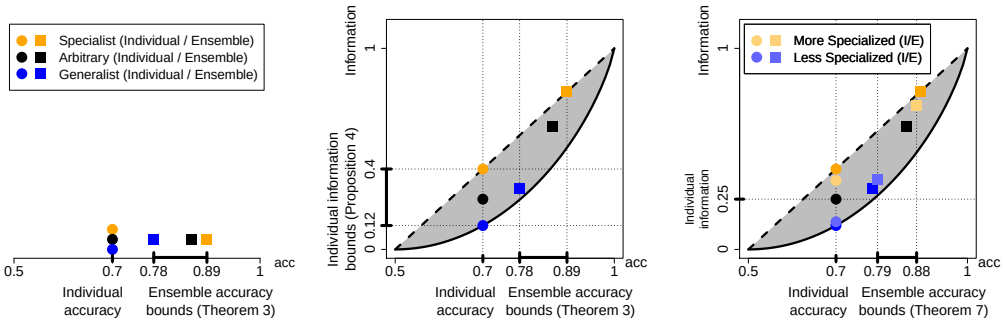

Figure 4.4: **Bounding ensemble accuracy with Theorem 3 and 7.** Left: Consider an individual classifier with an arbitrary and unknown confidence distribution f. We depict its known individual accuracy by a black dot (in this example, $\text{acc}_f = 70\%$). We construct a corresponding specialist and generalist classifier to this accuracy; their accuracies are depicted with blue and orange dots. Together, $k = 3$ arbitrary classifier with $\text{acc}_f = 70\%$ form an ensemble (with ensemble accuracy depicted by the black square). This ensemble accuracy is lower bounded by the accuracy of an ensemble of generalists (blue square) and upper bounded by that of an ensemble of specialists (orange square), see Theorem 3. Middle: In addition to the accuracy, we now also plot the information of classifiers, leading to a two-dimensional plot with accuracy on the x-axis and information on the y-axis. According to Proposition 4, the information of classifiers with accuracy acc_f lies between the information of the generalist and specialist. Generalists lie on the lower solid line and specialists lie on the upper dashed line. Thus, any arbitrary classifier's accuracy-information pair, $(\text{acc}_f, \text{I}_f)$, lies in the grey crescent shape. Again we depict the three individual classifiers of the left figure by dots and the corresponding ensemble classifiers by squares. Right: When the individual classifiers' information is known (here, 0.25 bit), Theorem 7 provides better bounds. Less specialized (light blue dot) and more specialized classifiers (light orange dot) form ensembles (same colored squares) whose accuracy bounds the ensemble accuracy of arbitrary classifiers. In this example, bounds from Theorem 3 improve only slightly but see Figure 4.3.

4.5.2 Improved Ensemble Accuracy Bounds

We will now assume that we know both, the accuracy and the information of the individual classifiers. Given these two measures, we can provide better bounds on the ensemble accuracy. These two measures still do not uniquely determine the confidence distribution of a classifier so that different ensemble accuracies are possible. As before (with specialists and generalists), we construct two confidence distributions: the more specialized classifier and the less specialized classifier. They will provide the new bounds.

The more specialized classifier (to a given accuracy acc and information I) is a mixture of specialist and generalist producing confidences at $C = 0.5$, $C = \text{acc}$ and $C = 1$, see Figure 4.2 (bottom left). The weights are such that the more specialized classifier can be shown to improve the ensemble accuracy.

Definition 5. ***(More specialized classifier)*** *A binary black-box classifier to the accuracy acc and information I is called more specialized if its confidence distribution is given by*

$$f^{\uparrow}_{acc,I} = w_{0.5}\delta_{0.5} + w_{acc}\delta_{acc} + w_1\delta_1,$$

with constants $w_{0.5} = \frac{2(1-acc)(I+g-1+H_2(acc))}{2acc-2+H_2(acc)}$, $w_{acc} = \frac{2acc-1-(I+g)}{2acc-2+H_2(acc)}$ *and* $w_1 = \frac{2(acc-0.5)(I+g-1+H_2(acc))}{2acc-2+H_2(acc)}$. *Constant g is defined in Section 4.6.3.*

The less specialized classifier is similar to a generalist but it can distinguish between slightly below average ($C = c^{\text{l}}$) and slightly above average confidences ($C = c^{\text{r}}$), see Figure 4.2 (bottom right).

Definition 6. ***(Less specialized classifier)*** *A binary black-box classifier to the accuracy acc and information I is called more specialized if its confidence distribution is given by*

$$f^{\downarrow}_{acc,I} = w_{c^l}\delta_{c^l} + w_{c^r}\delta_{c^r},$$

with constants $c^l = \frac{2(acc-0.5)(acc-I)-(1-acc)(1-H_2(acc))}{2(acc-0.5)(1-I)-2(1-acc)(1-H_2(acc))}$,
$c^r = \frac{2(acc-0.5)(acc-1+H_2(acc))-(1-acc)I}{2(acc-0.5)H_2(acc)-2(1-acc)I}$, *as well as* $w_{c^l} = \frac{c^r-acc}{c^r-c^l}$ *and* $w_{c^r} = \frac{acc-c^l}{c^r-c^l}$.

We can bound the ensemble accuracy of classifiers with known accuracies and information by the ensemble accuracies of more resp. less specialized classifiers.

Theorem 7. ***(More and less specialized classifiers bound the ensemble accuracy)*** *Consider k classifiers with individual accuracies* acc_i, *individual information* I_i *and confidence distributions* f_i *(*$i \in \{1..k\}$*). For each i,*

let $f^{\downarrow}_{acc_i,I_i}$ *and* $f^{\uparrow}_{acc_i,I_i}$ *be the less resp. more specialized classifier constructed to the accuracy and information of classifier* i*. Now consider the ensemble classifier based on the original classifiers with ensemble confidence distribution* $f_e = \bigotimes_{i=1}^{k} f_i$ *according to lCWMV as well as the ensemble of less and more specialized classifiers with ensemble confidence distributions* $f_e^{\downarrow} = \bigotimes_{i=1}^{k} f^{\downarrow}_{acc_i,I_i}$ *and* $f_e^{\uparrow} = \bigotimes_{i=1}^{k} f^{\uparrow}_{acc_i,I_i}$*. Then the accuracy of the original ensemble is lower and upper bounded by the accuracies of the less and more specialized ensembles:*

$$acc_{f_e^{generalist}} \leq acc_{f_e^{\downarrow}} \leq acc_{f_e} \leq acc_{f_e^{\uparrow}} \leq acc_{f_e^{specialist}}.$$

The proof is in Section 4.6.3. The proof idea is visualized in Figure 4.4. Theorem 7 shows that additionally knowing the information of the individual classifiers allows to predict the ensemble performance better than when only their accuracy is known, see Figure 4.3. In the retina example, if we know that the $k = 3$ classifiers in the ensemble have an accuracy of $\mathrm{acc}_1 = \mathrm{acc}_2 = \mathrm{acc}_3 = 70\%$ and also know that they provide in expectation $\mathrm{I}_1 = \mathrm{I}_2 = \mathrm{I}_3 = 0.36$ bit of information, we can improve the ensemble accuracy bounds from 78%–89% (with only known accuracies) to 85%–89%. This corresponds to Figure 4.3, second plot, orange shaded area for $k = 3$. A lower information of 0.15 bit would lead to bounds of 78%–84% (blue shaded area). While the bounds in Theorem 3 are tight, we do not know whether the bounds in Theorem 7 are tight.

One application of Theorem 7 is to determine how many classifiers are at least necessary to guarantee a target ensemble accuracy of, say, 95%, see Figure 4.5. If the individual classifiers have an accuracy of 70% and a high information of 0.36 bit (light orange line) an ensemble size of $k = 7$ classifiers is required. If their accuracy is the same but their information is lower (0.26 bit, light blue line), then $k = 13$ classifiers are required to achieve the target ensemble accuracy.

4.5.3 Bounds for the Ensemble Mutual Information

Up to now, we have bounded the ensemble accuracy, acc_{f_e}. Since we introduced the information of an individual classifier as a second measure, we can also consider the bounds for the ensemble information, that is, the mutual information between the true label and the ensemble's output, I_{f_e}. The ensemble information behaves much like the ensemble accuracy and is bounded by the same confidence distributions as before.

Proposition 8. ***(Specialists and generalists bound the ensemble information)*** *Consider* k *classifiers with individual accuracies* acc_i *and confidence distributions* f_i *(*$i \in \{1..k\}$*). For each* i*, let* $f^{generalist}_{acc_i}$*,* $f^{\downarrow}_{acc_i,I_i}$*,* $f^{\uparrow}_{acc_i,I_i}$ *and*

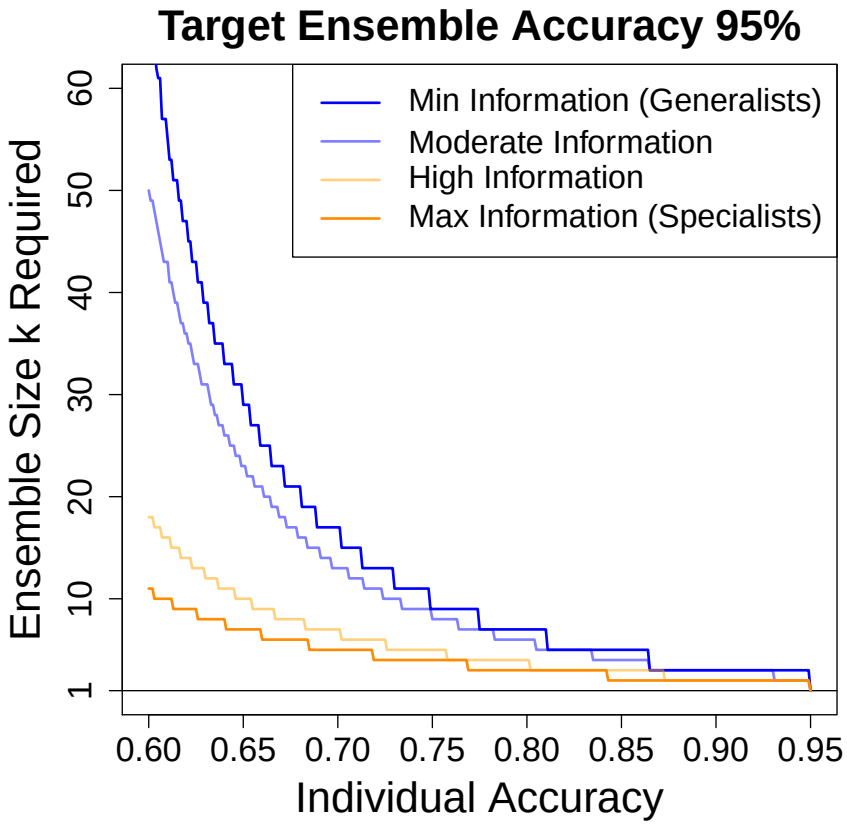

Figure 4.5: **Ensemble size determination.** To achieve a target ensemble accuracy of $\text{acc}_{f_e} = 95\%$ a certain ensemble size k (y-axis) is required depending on the accuracy of the individual classifiers (y-axis). Across accuracies, we consider 4 different levels of individual classifiers' information: minimal (blue), moderate (50% of the admissible information range; light blue), high (90%; light orange) and maximal (orange) information. Low information classifiers (blue) require larger ensembles to reach the target ensemble accuracy than high information classifiers (orange) with the same individual accuracy.

$f_{acc_i}^{specialist}$ *be as defined above. Now consider the ensemble classifier based on the original classifiers, with ensemble confidence distribution* $f_e = \bigotimes_{i=1}^{k} f_i$ *as well as ensembles with confidence distributions* $f_e^{generalist}$, $f_e^{\downarrow}$, $f_e^{\uparrow}$, *and* $f_e^{specialist}$ *as in Theorem 7. The information of the ensemble classifier is bounded by*

$$I_{f_e^{generalist}} \leq I_{f_e^{\downarrow}} \leq I_{f_e} \leq I_{f_e^{\uparrow}} \leq I_{f_e^{specialist}}.$$

The proof is in Section 4.6.4. Classifiers with confidence distributions that improve the ensemble accuracy also tend to improve the ensemble informa-

tion. Finally, we bound the ensemble information for when only the individual classifiers' information is known (but not their accuracies).

Proposition 9. ***(Information constrained specialists and generalists bound the ensemble information)*** *Consider k classifiers with individual information I_i and confidence distributions f_i ($i \in \{1..k\}$). For each i, let the accuracies corresponding to the individual information be $\tilde{acc}_i = H_2^{-1}(1 - I_i)$. Let $f^{generalist}_{\underset{\sim}{acc}_i}$ and $f^{specialist}_{\underset{\sim}{acc}_i}$ as defined above. Now consider the ensemble information based on the original classifiers, with ensemble confidence distribution $f_e = \bigotimes_{i=1}^{k} f_i$ as well as ensembles with confidence distributions $\tilde{f}_e^{generalist} = \bigotimes_{i=1}^{k} f^{generalist}_{\underset{\sim}{acc}_i}$ and $\tilde{f}_e^{specialist} = \bigotimes_{i=1}^{k} f^{specialist}_{\underset{\sim}{acc}_i}$ as in Theorem 7. The information of the ensemble classifier is bounded by*

$$I_{\tilde{f}_e^{generalist}} \leq I_{f_e} \leq I_{\tilde{f}_e^{specialist}}.$$

The proof is in Section 4.6.4. At first sight, this seems to be unsurprising: Again, specialists and generalists attain the upper resp. lower bounds. But specialists have a lower accuracy than generalists to the same information. Consider individual specialists with known information of $I_i = 0.4$ bit: In Figure 4.4 (middle) they lie on the dashed line to the left of the crescent (orange dot, with an accuracy of 70%). Generalists with the same information lie on the solid curve to the right (with accuracy of around 85%, no dot is shown). Having the same information, specialists have a 15%-point lower accuracy than generalists but nevertheless produce a better ensemble information! The explanation for this can be found by applying information decomposition (Griffith & Koch, 2014; Harder, Salge, & Polani, 2013): Specialists in our setting have a higher proportion of *unique* and a smaller proportion of *redundant* information as compared to generalists and are therefore more effective in ensembles. Thus, Proposition 9 demonstrates another desirable property of specialists.

4.6 Proofs

In this section, we show how two individual classifiers with confidence distributions f_1 and f_2 combine into an ensemble classifier with confidence distribution $f_e = f_1 \otimes f_2$. Throughout the proofs (except for Section 4.6.1 to remain consistent with the main text), we will consider only discrete confidence distributions $f(c) = P(C = c)$ with support $\Omega_f = \{c | f(c) > 0\}$ to keep notation simple.

To further simplify notation, we introduce an ad-hoc notation $f^*(C)$ to a given confidence distribution $f(C)$. It redistributes confidence mass from the range of $C \in [0.5, 1]$ to $C^* \in [0, 1]$ symmetrically: Half the probability mass of $f(C)$ goes to $f^*(C)$ and the other half to $f^*(1 - C)$.

Definition S1. ***(Redistributed confidence distribution)*** *Let $f : [0.5, 1] \to \mathbb{R}$ be a confidence distribution. Then $f^* : [0, 1] \to \mathbb{R}$ is the corresponding redistributed confidence distribution such that*

$$f^*(c) = \begin{cases} f(1-c)/2 & 0 \leq c < 0.5 \\ f(c) & c = 0.5 \\ f(c)/2 & 0.5 < c \leq 1 \end{cases} \quad \textit{and} \quad f(c) = \begin{cases} f^*(c) & c = 0.5 \\ 2f^*(c) & 0.5 < c \leq 1 \end{cases}$$

Let $g : (0, 1) \times (0, 1) \to (0, 1)$ be the function that determines which redistributed confidence, $c_2^* \in (0, 1)$, the second classifier has to produce such that together with the confidence of the first classifier, $c_1^* \in (0, 1)$, a given ensemble confidence, $c_e^* \in (0, 1)$, is obtained, $c_2^* = g(c_e^*, c_1^*)$. Then, the confidence distribution of the ensemble is given by Proposition S2.

Proposition S2. ***(Combining confidence distributions)*** *Given are two classifiers with confidence distributions f_1 and f_2. The ensemble confidence distribution $f_e = f_1 \otimes f_2$ is*

$$f_e(c_e) = \big(f_1 \otimes f_2\big)(c_e) = \begin{cases} \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} \left(f_1^*(c_1^*) \cdot f_2^*(g(c_e, c_1^*)) \cdot \frac{2c_1^*(1-c_1^*)}{c_1^* + c_e - 2c_1^* c_e} \right) \\ \qquad\qquad \textit{for } c_e = 0.5 \\ 2 \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} \left(f_1^*(c_1^*) \cdot f_2^*(g(c_e, c_1^*)) \cdot \frac{2c_1^*(1-c_1^*)}{c_1^* + c_e - 2c_1^* c_e} \right) \\ \qquad\qquad \textit{for } 0.5 < c_e < 1 \\ f_1(1) + f_2(1) - f_1(1) f_2(1) \\ \qquad\qquad \textit{for } c_e = 1 \end{cases}$$

where $g(c_e, c_1^*) = \frac{c_e(1-c_1^*)}{-2c_1^* c_e + c_1^* + c_e}$.

Proof. First, we show that $c_1^* \in (0, 1)$ together with $c_2^* = g(c_e^*, c_1^*) \in (0, 1)$ produces $c_e^* \in (0, 1)$. We rearrange

$$c_e^* = \frac{1}{1 + \exp\left(- \left(\log\left(\frac{c_1^*}{1-c_1^*} \right) + \log\left(\frac{c_2^*}{1-c_2^*} \right) \right) \right)}$$

so that

$$c_2^* = \frac{1}{1 + \exp\left(- \left(\log\left(\frac{c_e^*}{1-c_e^*} \right) - \log\left(\frac{c_1^*}{1-c_1^*} \right) \right) \right)} = \frac{c_e^*(1 - c_1^*)}{-2c_1^* c_e^* + c_1^* + c_e^*} = g(c_e^*, c_1^*).$$

We now show for the two cases, $c_e \in [0.5, 1)$ and $c_e = 1$, that the ensemble confidence distribution returns the probability that the ensemble prediction is correct.

(1) Case $c_e \in (0.5, 1)$: We remove $c_1^* = 0$ and $c_1^* = 1$ from the support Ω_{f^*} because, by convention, they produce $c_e^* = 0$ and $c_e^* = 1$ and therefore $c_e = 1$, which is excluded in this case.

$$
\begin{aligned}
f_e^*(c_e^*) &= \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} f_1^*(c_1^*) \cdot f_2^*(g(c_e^*, c_1^*)) \cdot \frac{2c_1^*(1-c_1^*)}{-2c_1^* c_e + c_1^* + c_e} \\
&= \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} f_1^*(c_1^*) \cdot f_2^*(g(c_e^*, c_1^*)) \cdot 2\Bigg(\frac{c_e^* c_1^*(1-c_1^*)}{-2c_1^* c_e + c_1^* + c_e} + \\
&\qquad\qquad \frac{(1-c_e^*)c_1^*(1-c_1^*)}{-2c_1^* c_e + c_1^* + c_e} \Bigg) \\
&= \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} f_1^*(c_1^*) \cdot f_2^*(g(c_e^*, c_1^*)) \cdot 2(c_1^* g(c_e^*, c_1^*) + \\
&\qquad\qquad (1-c_1^*)(g(1-c_e^*, 1-c_1^*))
\end{aligned}
$$

We use the symmetry, $g(1-c_e^*, 1-c_1^*) = 1 - g(c_e^*, c_1^*))$. Plugging these values into Definition S1 yields the next step.

$$
\begin{aligned}
f_e^*(c_e^*) &= \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} f_1^*(c_1^*) \cdot f_2^*(g(c_e^*, c_1^*)) \cdot 2(c_1^* g(c_e^*, c_1^*) + (1-c_1^*)(1 - g(c_e^*, c_1^*)) \\
&= \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} 2\Big(c_1^* \cdot f_1^*(c_1^*) \cdot g(c_e^*, c_1^*) \cdot f_2^*(g(c_e^*, c_1^*)) + \\
&\qquad\qquad (1-c_1^*) \cdot f_1^*(c_1^*) \cdot (1 - g(c_e^*, c_1^*)) \cdot f_2^*(g(c_e^*, c_1^*)) \Big) \\
&= \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} \sum_{y \in \{-1,+1\}} \Bigg(P(Y=y) \cdot \frac{P(Y=y|C_1^*=c_1^*)P(C_1^*=c_1^*)}{P(Y=y)} \cdot \\
&\qquad\qquad \frac{P(Y=y|C_2^*=g(c_e^*,c_1^*))P(C_2^*=g(c_e^*,c_1^*))}{P(Y=y)} \Bigg) \\
&= \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} \sum_{y \in \{-1,+1\}} P(Y=y) \cdot P(C_1^*=c_1^*|Y=y) \cdot \\
&\qquad\qquad P(C_2^*=g(c_e^*,c_1^*)|Y=y)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} \sum_{y \in \{-1,+1\}} P(C_1^* = c_1^*, C_2^* = g(c_e^*, c_1^*), Y = y) \\
&= \sum_{c_1^* \in \Omega_{f^*} \setminus \{0,1\}} P(C_1^* = c_1^*, C_2^* = g(c_e^*, c_1^*)) \\
&= P(C_e^* = c_e^*)
\end{aligned}
$$

(2) Case $c_e = 1$: We solve the edge case using the convention, $c_e = 1 \iff c_1 = 1$ or $c_2 = 1$. Then

$$
\begin{aligned}
f_e(c_e) &= f_1(1) + f_2(1) - f_1(1) f_2(1) \\
&= P(C_1 = 1) + P(C_2 = 1) - P(C_1 = 1 \wedge C_2 = 1) \\
&= P(C_1 = 1 \vee C_2 = 1) \\
&= P(C_e = 1).
\end{aligned}
$$

□

The operator $\otimes$ is closed on the space of confidence distributions (probability distributions over $C \in [0.5, 1]$). Its associativity and commutativity follow from associativity and commutativity of addition and multiplication. The neutral element is $f_{0.5}^{\text{generalist}}$. Together, this makes the operator $\otimes$ a commutative monoid.

4.6.1 Mutual Information Between True Label and Classifier Output

Here, we show that the mutual information between true label and a classifier's output is a function that only depends on the classifier's confidence distribution.

Proposition S3. ***(Information is a function of local confidences)*** *Given is a classifier as defined in the main paper that produces the predictions and confidences as output, $O = (\hat{Y}, C)$. The mutual information I between the true label Y and the classifier's output O is*

$$
I(Y; O) = \int_{0.5}^{1} f(c) \left(1 - H_2(c)\right) \, dc.
$$

Proof.

$$
\begin{aligned}
I(Y; O) &= H(Y) - H(Y|O) = H(Y) - H\left(Y|(\hat{Y}, C)\right) \\
&= \int_{0.5}^{1} f(c) \left(H(Y) - H(Y|\hat{Y}, C = c)\right) \, dc
\end{aligned}
$$

By assumption, Y is binary and equally weighted so that $H(Y) = H_2(0.5) = 1$ bit. To complete the proof, we have to show that $H(Y|\hat{Y}, C = c) = H_2(c)$:

$$
\begin{aligned}
&H(Y|\hat{Y}, C = c) \\
&= -\sum_{Y\in\{\pm1\}}\sum_{\hat{Y}\in\{\pm1\}} P(Y,\hat{Y}|C = c)\log_2 P(Y|\hat{Y}, C = c) \\
&= -\sum_{Y\in\{\pm1\}}\sum_{\hat{Y}\in\{\pm1\}} P(Y)P(\hat{Y}|Y, C = c)\log_2 P(Y|\hat{Y}, C = c) \\
&= -\sum_{Y\in\{\pm1\}}\sum_{\hat{Y}\in\{\pm1\}} \frac{1}{2}P(\hat{Y}|Y, C = c)\log_2 P(Y|\hat{Y}, C = c) \\
&= -\frac{1}{2}\left(\underbrace{c\log_2 c}_{Y=\hat{Y}=+1} + \underbrace{c\log_2 c}_{Y=\hat{Y}=-1} + \underbrace{(1-c)\log_2(1-c)}_{Y=+1\neq\hat{Y}=-1} + \underbrace{(1-c)\log_2(1-c)}_{Y=-1\neq\hat{Y}=+1}\right) \\
&= -\left(c\log_2 c + (1-c)\log_2(1-c)\right) \\
&= H_2(c).
\end{aligned}
$$

□

4.6.2 Refinement and Jensen's Inequality

For all remaining proofs, we will use a partial ordering on the classifiers, called refinement (DeGroot & Fienberg, 1983). In general, we will show here that more refined classifiers have higher scores on so called scoring functions. The remaining sections of the proofs then only aim to show that certain functions (the ensemble accuracy, the individual information etc.) are a convex scoring function.

Intuitively, we say a classifier with confidence distribution f is more refined than a classifier with confidence distribution f' if f', instead of producing different confidences c_1 and c_2, produces an intermediate confidence $c^{\text{center}} = tc_1 + (1-t)c_2$.

Definition S4. ***(Refinement)*** *A classifier with confidence distribution f is more refined than a classifier f', $f \succ f'$, if there exist c_1, $c_2 \in [0.5, 1]$, and*

$\epsilon_1, \epsilon_2 \in \mathbb{R}$ *such that* $0 \leq \epsilon_1 \leq f(c_1)$, $0 \leq \epsilon_2 \leq f(c_2)$ *and*

$$f'(c) = \begin{cases} f(c) & c \neq c_1, c \neq c_2, c \neq c^{center} \\ f(c) - \epsilon_1 & c = c_1 \\ f(c) - \epsilon_2 & c = c_2 \\ f(c) + \epsilon_1 + \epsilon_2 & c = c^{center}. \end{cases}$$

where c^{center} *is the weighted mean,* $c^{center} = \frac{\epsilon_1 c_1 + \epsilon_2 c_2}{\epsilon_1 + \epsilon_2}$. *Furthermore, if* $f \succ f'$ *and* $f' \succ f''$ *then* $f \succ f''$ *(transitivity).*

In the main paper, we have considered four particular classifiers: Specialist, more specialized classifier, less specialized classifier and generalist. For a given accuracy, acc, and information, I, these classifiers are in a refinement ordering, $f_{\text{acc}}^{\text{specialist}} \succ f_{\text{acc,I}}^{\uparrow} \succ f_{\text{acc,I}}^{\downarrow} \succ f_{\text{acc,I}}^{\text{generalist}}$. For example, it is straight forward to see that a specialist is more refined than a generalist by choosing $c_1 = 0.5$, $c_2 = 1$, $\epsilon_1 = w_{0.5}$ and $\epsilon_2 = w_1$ in Definition S4 to obtain $f_{\text{acc}}^{\text{generalist}}$.

We evaluate confidence distributions, for example, by computing the accuracy or information. These evaluations are based on scoring functions, $\phi : [0.5, 1] \to \mathbb{R}$, that translate local confidences into values $\phi(c)$.

Definition S5. ***(Score)*** *Given a scoring function* $\phi(c) : [0.5, 1] \to \mathbb{R}$*, the score of a confidence distribution* f *is*

$$\Phi(f) = \sum_{c \in \Omega_c} f(c)\phi(c)\ dc.$$

For example, when we choose the scoring function $\phi(c) = 1 - H_2(c)$ to evaluate a classifier's confidence distribution f, the score is the information, $\Phi(f) = \text{I}_f = \int_{0.5}^{1} f(c)(1 - H_2(c))\ dc$. When we chose the identity scoring function $\phi(c) = c$, the score is the accuracy $\Phi(f) = \text{acc}_f = \int_{0.5}^{1} f(c)c\ dc$.

For convex scoring functions, we can apply apply Jensen's inequality,

$$\phi(tc_1 + (1-t)c_2) \leq t\phi(c_1) + (1-t)\phi(c_2),$$

to show that less refined confidence distributions (generalists) produce lower scores while more refined confidence distributions (specialists) produce higher scores.

Lemma S6. ***(Jensen's inequality for confidence distributions)*** *Let* ϕ *be a convex scoring function with score* $\Phi_f = \sum_{c \in \Omega_f} f(c)\phi(c)$. *If* f *is more refined than* f' *then* $acc_f = acc_{f'}$ *and* $\Phi_f \geq \Phi'_f$.

Proof. First, f' has the same accuracy as the original f:

$$\begin{aligned}
\text{acc}_{f'} &= \text{acc}_f \\
\sum_{c \in \Omega'_f} f'(c)c &= \sum_{c \in \Omega_f} f(c)c \\
f'(c^{\text{center}})c^{\text{center}} &= \epsilon_1 c_1 + \epsilon_2 c_2 + f(c^{\text{center}})c^{\text{center}} \\
f'(c^{\text{center}})c^{\text{center}} &= \epsilon_1 c_1 + \epsilon_2 c_2 + f(c^{\text{center}})c^{\text{center}} \\
f'(c^{\text{center}})c^{\text{center}} &= (\epsilon_1 + \epsilon_2)\frac{\epsilon_1 c_1 + \epsilon_2 c_2}{\epsilon_1 + \epsilon_2} + f(c^{\text{center}})c^{\text{center}} \\
f'(c^{\text{center}})c^{\text{center}} &= (\epsilon_1 + \epsilon_2)c^{\text{center}} + f(c^{\text{center}})c^{\text{center}} \\
f'(c^{\text{center}})c^{\text{center}} &= (\epsilon_1 + \epsilon_2 + f(c^{\text{center}}))c^{\text{center}} \\
f'(c^{\text{center}})c^{\text{center}} &= f'(c^{\text{center}})c^{\text{center}}.
\end{aligned}$$

Second, f' has a smaller (or equal) score Φ as f.

$$\begin{aligned}
\Phi'_f &\leq \Phi_f \\
\sum_{c \in \Omega'_f} f'(c)\phi(c) &\leq \sum_{c \in \Omega_f} f(c)\phi(c) \\
f'(c^{\text{center}})\phi(c^{\text{center}}) &\leq \epsilon_1 \phi(c_1) + \epsilon_2 \phi(c_2) + f(c^{\text{center}})\phi(c^{\text{center}}) \\
(f(c^{\text{center}}) + \epsilon_1 + \epsilon_2)\phi(c^{\text{center}}) &\leq (\epsilon_1 + \epsilon_2)\frac{\epsilon_1 \phi(c_1) + \epsilon_2 \phi(c_2)}{\epsilon_1 + \epsilon_2} + f(c^{\text{center}})\phi(c^{\text{center}}) \\
\phi(c^{\text{center}}) &\leq \frac{\epsilon_1 \phi(c_1) + \epsilon_2 \phi(c_2)}{\epsilon_1 + \epsilon_2} \\
\phi\left(\frac{\epsilon_1 c_1 + \epsilon_2 c_2}{\epsilon_1 + \epsilon_2}\right) &\leq \frac{\epsilon_1 \phi(c_1) + \epsilon_1 f(c_2)\phi(c_2)}{\epsilon_1 + \epsilon_2} \\
\phi\left(t c_1 + (1-t)c_2\right) &\leq t\phi(c_1) + (1-t)\phi(c_2)
\end{aligned}$$

The last inequality holds due to Jensen's inequality for convex ϕ. □

The immediate consequence is that generalists produce the lowest score and specialists produce the highest score for any convex scoring function.

Corollary S7. ***(Generalist and specialist produce minimal and maximal value of convex scoring functions)*** *Let $f(c)$ be a confidence distribution with fixed accuracy acc_f and $\phi(c)$ be a convex scoring function with score Φ_f. The score is minimized by the generalist and maximized by the specialist.*

$$\min_f \Phi_f = \Phi_f^{generalist} \qquad \textit{and} \qquad \max_f \Phi_f = \Phi_f^{specialist}$$

Proof. For any confidence distribution f with accuracy $\text{acc} = \text{acc}_f$ and probability mass at different confidences, $c_1 \neq c_2$ with $f(c_1) > 0$ and $f(c_2) > 0$, $\exists : f' : f' \prec f$ so that $\Phi(f') \leq \Phi(f)$. By induction, $\min_f \Phi_f$ is obtained by the least refined confidence distribution, $f_{\text{acc}}^{\text{generalist}}$.

For any confidence distribution f' with accuracy $\text{acc} = \text{acc}_{f'}$ and probability mass at confidence $0.5 < c^{\text{center}} < 1$ with $f(c^{\text{center}}) > 0$, $\exists f : f' \prec f$ such that $\Phi(f') \leq \Phi(f)$. By induction, $\max_f \Phi_f$ is obtained by the most refined confidence distribution, $f_{\text{acc}}^{\text{specialist}}$. □

4.6.3 Ensemble Accuracy Bounds

We now prove the bounds on the ensemble accuracy. Confidence distributions f in the following proofs will be discrete probability distributions with support $\Omega_f = \{c|f(c) > 0\}$ to simplify notation. The continuous case follows by generalizing Jensen's inequality to the continuous functions. The tricky part of the proofs is not handling the continuous case; the tricky part is to show convexity of several functions so that we can apply Corollary S7.

Theorem 3. ***(Specialists and generalists bound the ensemble accuracy)*** *Consider k classifiers with individual accuracies acc_i and confidence distributions f_i ($i \in \{1..k\}$). For each i, let $f_{acc_i}^{generalist}$ and $f_{acc_i}^{specialist}$ be a generalist resp. specialist classifier that has the same accuracy as classifier i. Now consider the ensemble classifier based on the original classifiers with ensemble confidence distribution $f_e = \bigotimes_{i=1}^{k} f_i$ according to lCWMV as well as the ensemble of generalists and ensemble of specialists with ensemble confidence distributions $f_e^{generalist} = \bigotimes_{i=1}^{k} f_{acc_i}^{generalist}$ and $f_e^{specialist} = \bigotimes_{i=1}^{k} f_{acc_i}^{specialist}$. Then the accuracy of the original ensemble is lower and upper bounded by the accuracies of the generalist and specialist ensembles:*

$$acc_{f_e^{generalist}} \leq acc_{f_e} \leq acc_{f_e^{specialist}} \ .$$

Proof. The ensemble accuracy for two classifiers is

$$\begin{aligned} \text{acc}_{f_e} &= \sum_{c \in \Omega_{f_e}} f_e(c)c \\ &= \sum_{c_1 \in \Omega_{f_1}} \sum_{c_2 \in \Omega_{f_2}} f_1(c_1) f_2(c_2) P(\hat{y}_e \text{ correct}|c_1, c_2). \end{aligned}$$

Expanding $P(\hat{y}_e \text{ correct}|c_1, c_2)$ yields

$$\begin{aligned} P(\hat{y}_e \text{ correct}|c_1, c_2) &= P(\hat{y}_e \text{ correct}, \hat{y}_1 = \hat{y}_2|c_1, c_2) + P(\hat{y}_e \text{ correct}, \hat{y}_1 \neq \hat{y}_2|c_1, c_2) \\ &= c_1 c_2 + \max\{c_1(1-c_2), (1-c_1)c_2\} \\ &= \max\{c_1, c_2\} \end{aligned}$$

In continuation, the ensemble accuracy is

$$\text{acc}_{f_e} = \sum_{c_1 \in \Omega_{f_1}} \sum_{c_2 \in \Omega_{f_2}} f_1(c_1) f_2(c_2) \max\{c_1, c_2\}.$$

The function $\phi(c_1) = \sum_{c_2 \in \Omega_{f_2}} f_2(c_2) \max\{c_1, c_2\}$ is convex in c_1 for any c_2 because max is convex. The sum (over c_2) of convex functions remains convex. Thus, $\phi(c_1)$ is a convex scoring function for f_1. Corollary S7 yields that generalists vs. specialists minimize vs. maximize the score, proving the desired statement for two classifiers. By induction, we obtain the desired result.

□

Since not only the ensemble accuracy is convex in individual confidences but also the mutual information, generalists and specialists yield minimal and maximal values in both cases.

Proposition S8. ***(Specialists and generalists bounds the individual information)*** *A classifier with confidence distribution f and accuracy acc has an information between*

$$I_{f_{acc}^{generalist}} \leq I_f \leq I_{f_{acc}^{specialist}}.$$

Proof. The information is

$$\mathrm{I}_f = I(Y; (\hat{Y}, C)) = \sum_{c \in \Omega_f} f(c) \left(1 - H_2(c)\right).$$

We derive the second derivative.

$$\begin{aligned} 1 - H_2(c) &= 1 - \left(c \log_2 \left(\frac{1}{c} \right) + (1-c) \log_2 \left(\frac{1}{1-c} \right) \right) \\ \frac{\mathrm{d}}{\mathrm{d}c} \left(1 - H_2(c)\right) &= \log_2 \left(\frac{c}{1-c} \right) \\ \frac{\mathrm{d}^2}{\mathrm{d}c^2} \left(1 - H_2(c)\right) &= \frac{1}{\log_e(2) c (1-c)} \end{aligned}$$

The second derivative is strictly larger than 0 in $c \in [0.5, 1)$ so that $\phi(c) = 1 - H_2(c)$ is convex. Corollary S7 yields the desired statement. □

Individual accuracy and information of a classifier still do not determine its confidence distribution. For the proof of Theorem 7, we follow this strategy: For each classifier f with given accuracy and information, we construct a more refined classifier $f^{\nearrow}$. Since there are multiple f that satisfy the two constraints, there are multiple $f^{\nearrow}$. We then construct one unique $f^{\uparrow}$ (more specialized classifier) that is more refined than any $f^{\nearrow}$. By transitivity, $f_i \prec f^{\nearrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}} \prec f^{\uparrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}}$. Therefore, $f^{\uparrow}$ improves the ensemble accuracy.

Analogously, f is less refined than $f^{\searrow}$ and the unique $f^{\downarrow}$ (less specialized classifier) is even less refined than $f^{\searrow}$. Thus $f^{\downarrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}} \prec f^{\searrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}} \prec f_i$ and $f^{\downarrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}}$ makes the ensemble accuracy worse.

We will consider the left conditional confidence distribution and the right conditional confidence distribution. By that we mean the confidence distribution $f(C)$ conditioned on $C < \mathrm{acc}_f$ resp. $C \geq \mathrm{acc}_f$. The probabilities to obtain a below or above average confidence are $p^{\mathrm{left}} = P(C < \mathrm{acc}_f)$ resp. $p^{\mathrm{right}} = P(C \geq \mathrm{acc}_f)$. The left and right conditional accuracies are

$$\mathrm{acc}_f^{\mathrm{left}} = \sum_{c\in\Omega_f, c<\mathrm{acc}_f} \frac{f(c)}{p^{\mathrm{left}}} \cdot c \qquad \text{and} \qquad \mathrm{acc}_f^{\mathrm{right}} = \sum_{c\in\Omega_f, c\geq\mathrm{acc}_f} \frac{f(c)}{p^{\mathrm{left}}} \cdot c.$$

These are the accuracies of the classifier f when only counting below average (left) or above average (right) confidences. Analogously, the left and right conditional information are

$$\mathrm{I}_f^{\mathrm{left}} = \sum_{c\in\Omega_f, c<\mathrm{acc}_f} \frac{f(c)}{p^{\mathrm{left}}} \cdot (1 - H_2(c)) \quad \text{and} \quad \mathrm{I}_f^{\mathrm{right}} = \sum_{c\in\Omega_f, c\geq\mathrm{acc}_f} \frac{f(c)}{p^{\mathrm{right}}} \cdot (1 - H_2(c)).$$

Now, we can define the more and the less refined classifier we need for the proof. The idea for the more refined classifier is to split confidences below average ($C < \mathrm{acc}_f$) to $C = 0.5$ and $C = \mathrm{acc}_f$ and to split confidences above average ($C \geq \mathrm{acc}_f$) to $C = \mathrm{acc}_f$ and $C = 1$. We end up with a classifier that outputs only three possible confidences: $C = 0.5$, $C = \mathrm{acc}_f$ and $C = 1$. The probability masses for these cases depend on the original classifier's confidence distribution f.

Definition S9. ***(More refined classifier)*** *A binary black-box classifier to the accuracy acc and information I is called more refined if its confidence distribution is given by*

$$f^{\nearrow}_{acc_f,I_f} = w^{\nearrow}_{0.5}\delta_{acc} + w^{\nearrow}_{acc_f}\delta_{acc_f} + w^{\nearrow}_{1}\delta_1$$

with constants $w^{\nearrow}_{0.5} = \frac{acc_f - acc_f^{left}}{acc_f - 0.5}$, $w^{\nearrow}_{acc_f} = \frac{acc_f^{left} - 0.5}{acc_f - 0.5} + \frac{1 - acc_f^{right}}{1 - acc_f}$, *and* $w^{\nearrow}_{1} = \frac{acc_f^{right} - acc_f}{1 - acc_f}$.

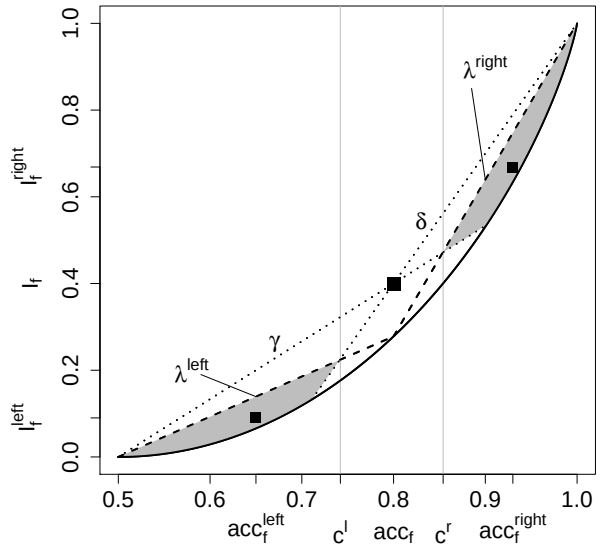

Figure 4.6: **Proof idea of Theorem 7** An individual classifier's confidence distribution f is characterized by a point in the accuracy-information plot. It has accuracy acc_f (x-axis coordinate) and information I_f (y-axis coordinate). Consider the left and right conditional distributions that conditioned on $C < \mathrm{acc}_f$ and $C \geq \mathrm{acc}_f$, respectively. These confidence distributions have the (accuracy, information)-pairs: ($\mathrm{acc}^{\mathrm{left}}$, $\mathrm{I}^{\mathrm{left}}$), and ($\mathrm{acc}^{\mathrm{right}}$, $\mathrm{I}^{\mathrm{right}}$), which must lie in the shaded areas.

In analogy, the less refined classifier does not split the left and right conditional confidence distributions but fully merges them into $C = \mathrm{acc}_f^{\mathrm{left}}$ and $C = \mathrm{acc}_f^{\mathrm{right}}$.

Definition S10. ***(Less refined classifier)*** *A binary black-box classifier to the accuracy acc and information I is called less refined if its confidence distribution is given by*

$$f_{acc,I}^{\searrow} = w_{acc_f^{left}}^{\searrow} \delta_{acc_f^{left}} + w_{acc_f^{right}}^{\searrow} \delta_{acc_f^{right}},$$

with constants $w_{acc_f^{left}}^{\searrow} = \frac{acc_f^{right} - acc_f}{acc_f^{right} - acc_f^{left}}$, *and* $w_{acc_f^{left}}^{\searrow} = \frac{acc_f - acc_f^{left}}{acc_f^{right} - acc_f^{left}}$.

These two classifiers, even though similar, are different to the more specialized resp. less specialized classifier. We exploit that they are in a refinement relation to them, which allows us to prove Theorem 7.

Theorem 7. ***(More and less specialized classifiers bound the ensemble accuracy)*** *Consider k classifiers with individual accuracies acc_i, individual information I_i and confidence distributions f_i ($i \in \{1..k\}$). For each i, let $f^{\downarrow}_{acc_i,I_i}$ and $f^{\uparrow}_{acc_i,I_i}$ be the less resp. more specialized classifier constructed to the accuracy and information of classifier i. Now consider the ensemble classifier based on the original classifiers with ensemble confidence distribution $f_e = \bigotimes_{i=1}^{k} f_i$ according to lCWMV as well as the ensemble of less and more specialized classifiers with ensemble confidence distributions $f_e^{\downarrow} = \bigotimes_{i=1}^{k} f^{\downarrow}_{acc_i,I_i}$ and $f_e^{\uparrow} = \bigotimes_{i=1}^{k} f^{\uparrow}_{acc_i,I_i}$. Then the accuracy of the original ensemble is lower and upper bounded by the accuracies of the less and more specialized ensembles:*

$$acc_{f_e^{generalist}} \leq acc_{f_e^{\downarrow}} \leq acc_{f_e} \leq acc_{f_e^{\uparrow}} \leq acc_{f_e^{specialist}}.$$

Proof. First, we show the upper bound in (1) and then the lower bound in (2). Our strategy will be to prove that there is a refinement ordering, $\forall i \in \{1..k\} : f^{\downarrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}} \prec f^{\searrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}} \prec f_i \prec f^{\nearrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}} \prec f^{\uparrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}}$. Because more refined classifiers also produce higher ensemble accuracies, Lemma S6 produces the desired result.

(1) By construction, $f_i \prec f^{\nearrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}}$. Let the information gain be $g_{f_i} = \mathrm{I}_{f^{\nearrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}}} - \mathrm{I}_{f_i}$. Let f^* be the confidence distribution that produces the maximal gain, $f^* = \mathrm{argmax}_f\, g_f$ s.t. $\mathrm{acc}_{f_i} = \mathrm{acc}_i$ and $\mathrm{I}_{f_i} = \mathrm{I}_i$. Let the maximal gain be $g = g_{f^*}$ (this is the constant in Definition 5). It remains to show that $\forall f : f^{\nearrow}_{\mathrm{acc}_f,\mathrm{I}_f} \prec f^{\uparrow}_{\mathrm{acc}_f,\mathrm{I}_f}$.

For all f, the more refined classifier distribution $f^{\nearrow}_{\mathrm{acc}_f,\mathrm{I}_f}$ is defined by constants $w^{\nearrow}_{0.5} = \frac{2(1-\mathrm{acc}_f)(\mathrm{I}_f+g_f-1+H_2(\mathrm{acc}_f))}{2\mathrm{acc}_f-2+H_2(\mathrm{acc}_f)}$, $w^{\nearrow}_{\mathrm{acc}_f} = \frac{2\mathrm{acc}_f-1-(\mathrm{I}_f+g_f)}{2\mathrm{acc}_f-2+H_2(\mathrm{acc}_f)}$ and $w^{\nearrow}_{1} = \frac{2(\mathrm{acc}_f-0.5)(\mathrm{I}_f+g_f-1+H_2(\mathrm{acc}_f))}{2\mathrm{acc}_f-2+H_2(\mathrm{acc}_f)}$. The more specialized classifier distribution is $f^{\nearrow}_{\mathrm{acc}_f,\mathrm{I}_f}$ with constants $w_{0.5}$, w_{acc_f} and w_1 as in Definition 5. To prove $f^{\nearrow}_{\mathrm{acc}_f,\mathrm{I}_f} \prec f^{\uparrow}_{\mathrm{acc}_f,\mathrm{I}_f}$, we apply Lemma S6 with $c_1 = 0.5$, $c_2 = 1$, $\epsilon_1 = w_{0.5} - w^{\nearrow}_{0.5}$, $\epsilon_2 = w_1 - w^{\nearrow}_{1}$. It remains to show that these constants transform the more specialized classifier into the more refined classifier: $\epsilon_1 + \epsilon_2 + w_{\mathrm{acc}_f} = w^{\nearrow}_{\mathrm{acc}_f}$ and that $c^{\mathrm{center}} = \mathrm{acc}_f$:

$$
\begin{aligned}
\epsilon_1 + \epsilon_2 + w_{\mathrm{acc}_f} &= w_{0.5} - w^{\nearrow}_{0.5} + w_1 - w^{\nearrow}_1 + w_{\mathrm{acc}_f} \\
&= 1 - w^{\nearrow}_{0.5} - w^{\nearrow}_1 \\
&= w^{\nearrow}_{\mathrm{acc}_f} \\
c^{\mathrm{center}} &= \frac{\epsilon_1 c_1 + \epsilon_2 c_2}{\epsilon_1 + \epsilon_2} \\
&= \frac{2(1-\mathrm{acc}_f)\cdot 0.5 + 2(\mathrm{acc}_f - 0.5)\cdot 1}{2(1-\mathrm{acc}_f) + 2(\mathrm{acc}_f - 0.5)} \\
&= 2(1-\mathrm{acc}_f)\cdot 0.5 + 2(\mathrm{acc}_f - 0.5) \\
&= \mathrm{acc}_f
\end{aligned}
$$

Taken together, $f_i \prec f^{\nearrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}} \prec f^{\uparrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}}$. By Lemma S6 follows the desired result for (1).

(2) By construction, $f^{\searrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}} \prec f_i$. The left conditional accuracy and information pair, $(\mathrm{acc}_f^{\mathrm{left}}, \mathrm{I}_f^{\mathrm{left}})$, must lie below the line λ^{left} that runs through points $(0.5, 0)$ and $(\pi, 1 - H_2(\pi))$ because of Proposition 4. See Figure 4.6 for a visualization. The right conditional information pair, $(\mathrm{acc}_f^{\mathrm{right}}, \mathrm{I}_f^{\mathrm{right}})$, must lie below the line λ^{right} that runs through $(\pi, 1 - H_2(\pi))$ and $(1, 1)$. In consequence, the left conditional pair must lie above the line δ running through $(\mathrm{acc}_f, \mathrm{I}_f)$ and $(1, 1)$: Assuming for the sake of contradiction that this was not the case entails that the right conditional pair would have to lie above λ^{right}. Analogously, the right conditional pair must lie below line γ that runs through $(0.5, 0)$ and $(\mathrm{acc}_f, \mathrm{I}_f)$. Thus, the innermost left and right conditional accuracies are the intersections of these lines, c^{l} resp. c^{r}, see Definition 6 and see the edges of the grey areas in Figure 4.6 that point towards the center.

For all f, the less refined classifier's confidence distribution is $f^{\searrow}_{\mathrm{acc}_f,\mathrm{I}_f}$ with left and right conditional accuracies $\mathrm{acc}_f^{\mathrm{left}}$ and $\mathrm{acc}_f^{\mathrm{right}}$. The less specialized classifier's confidence distribution is $f^{\downarrow}_{\mathrm{acc}_f,\mathrm{I}_f}$ with left and right conditional accuracies c^{l} and c^{r}. To prove $f^{\downarrow}_{\mathrm{acc}_f,\mathrm{I}_f} \prec f^{\searrow}_{\mathrm{acc}_f,\mathrm{I}_f}$, we apply Lemma S6 with constants $c_1 = \mathrm{acc}_f^{\mathrm{left}}$, $c_2 = \mathrm{acc}_f^{\mathrm{right}}$ twice: (a) with $\epsilon_1' = \frac{(\mathrm{acc}_f^{\mathrm{right}} - c^{\mathrm{l}})(c^{\mathrm{r}} - \mathrm{acc}_f)}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{right}} - c^{\mathrm{left}})}$ and $\epsilon_2' = \frac{(\mathrm{acc}_f^{\mathrm{left}} - c^{\mathrm{l}})(c^{\mathrm{r}} - \mathrm{acc}_f)}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{right}} - c^{\mathrm{left}})}$; and (b) with $\epsilon_1'' = \frac{(c^{\mathrm{r}} - \mathrm{acc}_f^{\mathrm{left}})(\mathrm{acc}_f - c^{\mathrm{l}})}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{right}} - c^{\mathrm{left}})}$ and $\epsilon_2'' = \frac{(c^{\mathrm{r}} - \mathrm{acc}_f^{\mathrm{left}})(\mathrm{acc}_f - c^{\mathrm{l}})}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{right}} - c^{\mathrm{left}})}$. It remains to show that with these constants transform the less refined classifier into the less specialized classifier: $\epsilon_1' + \epsilon_1'' = w_{\mathrm{acc}_f^{\mathrm{left}}}$, $\epsilon_2' + \epsilon_2'' = w_{\mathrm{acc}_f^{\mathrm{right}}}$, $\epsilon_1' + \epsilon_2' = w_{c^{\mathrm{l}}}$ and $\epsilon_1'' + \epsilon_2'' = w_{c^{\mathrm{r}}}$. $c^{\mathrm{center}\prime} = c^{\mathrm{l}}$, and $c^{\mathrm{center}\prime\prime} = c^{\mathrm{r}}$.

The two deviations on the left hand side deviations add up to the total weight of the left side of the less refined classifier.

$$\begin{aligned}
\epsilon_1' + \epsilon_1'' &= \frac{(\mathrm{acc}_f^{\mathrm{right}} - c^{\mathrm{l}})(c^{\mathrm{r}} - \mathrm{acc}_f)}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{r}} - c^{\mathrm{l}})}(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{r}} - c^{\mathrm{l}}) \\
&= \frac{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f)(c^{\mathrm{r}} - c^{\mathrm{l}})}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{r}} - c^{\mathrm{l}})} \\
&= \frac{\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f}{\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}}} \\
&= w_{\mathrm{acc}_f^{\mathrm{left}}}
\end{aligned}$$

The deviations produce the left weight of the less specialized classifier.

$$\begin{aligned}
\epsilon_1' + \epsilon_2' &= \frac{(\mathrm{acc}_f^{\mathrm{right}} - c^{\mathrm{l}})(c^{\mathrm{r}} - \mathrm{acc}_f)}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{right}} - c^{\mathrm{left}})} + \frac{(\mathrm{acc}_f^{\mathrm{left}} - c^{\mathrm{l}})(c^{\mathrm{r}} - \mathrm{acc}_f)}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{right}} - c^{\mathrm{left}})} \\
&= \frac{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{r}} - \mathrm{acc}_f)}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{right}} - c^{\mathrm{left}})} \\
&= \frac{c^{\mathrm{r}} - \mathrm{acc}_f}{c^{\mathrm{right}} - c^{\mathrm{left}}} \\
&= w_{c^{\mathrm{l}}}
\end{aligned}$$

The left hand side accuracy is kept constant.

$$\begin{aligned}
c^{\mathrm{center}\prime} &= \frac{\epsilon_1' c_1 + \epsilon_2' c_2}{\epsilon_1' + \epsilon_2'} \\
&= \frac{\frac{(\mathrm{acc}_f^{\mathrm{right}} - c^{\mathrm{l}})(c^{\mathrm{r}} - \mathrm{acc}_f)}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{r}} - c^{\mathrm{l}})}\mathrm{acc}_f^{\mathrm{left}} + \frac{(c^{\mathrm{l}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{r}} - \mathrm{acc}_f)}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{r}} - c^{\mathrm{l}})}\mathrm{acc}_f^{\mathrm{right}}}{\frac{(\mathrm{acc}_f^{\mathrm{right}} - c^{\mathrm{l}})(c^{\mathrm{r}} - \mathrm{acc}_f)}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{r}} - c^{\mathrm{l}})} + \frac{(c^{\mathrm{l}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{r}} - \mathrm{acc}_f)}{(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})(c^{\mathrm{r}} - c^{\mathrm{l}})}} \\
&= \frac{c^{\mathrm{l}}(\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}})}{\mathrm{acc}_f^{\mathrm{right}} - \mathrm{acc}_f^{\mathrm{left}}} \\
&= c^{\mathrm{l}}
\end{aligned}$$

The right side follows analogously. Taken together, $f^{\downarrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}} \prec f^{\searrow}_{\mathrm{acc}_{f_i},\mathrm{I}_{f_i}} \prec f_i$. By Lemma S6 follows the desired result (2).

□

4.6.4 Ensemble Information Bounds

Here, we proof the bounds on the ensemble information. To avoid clutter, we now use the natural logarithm $\log_e$ instead of $\log_2$ as in the main text and drop the subscript. Results are transferable because convexity does not change with the base of the logarithm.

Proposition 8. ***(Specialists and generalists bound the ensemble information)*** *Consider k classifiers with individual accuracies acc_i and confidence distributions f_i ($i \in \{1..k\}$). For each i, let $f_{acc_i}^{generalist}$, $f_{acc_i,I_i}^{\downarrow}$, $f_{acc_i,I_i}^{\uparrow}$ and $f_{acc_i}^{specialist}$ be as defined above. Now consider the ensemble classifier based on the original classifiers, with ensemble confidence distribution $f_e = \bigotimes_{i=1}^{k} f_i$ as well as ensembles with confidence distributions $f_e^{generalist}$, $f_e^{\downarrow}$, $f_e^{\uparrow}$, and $f_e^{specialist}$ as in Theorem 7. The information of the ensemble classifier is bounded by*

$$I_{f_e^{generalist}} \leq I_{f_e^{\downarrow}} \leq I_{f_e} \leq I_{f_e^{\uparrow}} \leq I_{f_e^{specialist}}.$$

Proof. The ensemble information is

$$\begin{aligned}
\mathrm{I}_{f_e} &= \sum_{c \in \Omega_{f_e}} f_e(c)\,(H_2(0.5) - H_2(c)) \\
&= \sum_{c_1 \in \Omega_{f_1}} \sum_{c_2 \in \Omega_{f_2}} f_1(c_1) f_2(c_2) \cdot \\
&\qquad \Big(P(\hat{y}_1 = \hat{y}_2 | c_1, c_2)(H_2(0.5) - H_2(P(\hat{y}_e \text{ correct} | c_1, c_2, \hat{y}_1 = \hat{y}_2))) \\
&\qquad + P(\hat{y}_1 \neq \hat{y}_2 | c_1, c_2)(H_2(0.5) - H_2(P(\hat{y}_e \text{ correct} | c_1, c_2, \hat{y}_1 \neq \hat{y}_2))) \Big)
\end{aligned}$$

We will use the following notation to simplify the term in the big brackets.

$$\begin{aligned}
\nu_{c_1,c_2}^{\text{agree}} &= c_1 c_2 + (1 - c_1)(1 - c_2) && \text{(conditional probability to agree)} \\
\nu_{c_1,c_2}^{\text{disagree}} &= c_1(1 - c_2) + c_1(1 - c_2) && \text{(conditional probability to disagree)} \\
\eta_{c_1,c_2}^{\text{agree}} &= \tfrac{c_1 \cdot c_2}{\nu^{\text{agree}}} && \text{(conditional confidence upon agreement)} \\
\eta_{c_1,c_2}^{\text{disagree}} &= \tfrac{\max\{c_1 \cdot (1-c_2), c_1 \cdot (1-c_2)\}}{\nu_{c_1,c_2}^{\text{disagree}}} && \text{(conditional confidence upon disagreement)}
\end{aligned}$$

With this notation, the term in the big brackets is

$$\phi_{c_2}(c_1) = \left(\nu^{\text{agree}}_{c_1,c_2}(H_2(0.5) - H_2(\eta^{\text{agree}}_{c_1,c_2})) + \nu^{\text{disagree}}_{c_1,c_2}(H_2(0.5) - H_2(\eta^{\text{disagree}}_{c_1,c_2}))\right)$$

and we will show that it is convex in c_1. We do so by showing that its second derivative is non-negative. First, we rearrange.

$$\begin{aligned}
\phi_{c_2}(c_1) &= \left(\nu^{\text{agree}}_{c_1,c_2}(H_2(0.5) - H_2(\eta^{\text{agree}}_{c_1,c_2})) + \nu^{\text{disagree}}_{c_1,c_2}(H_2(0.5) - H_2(\eta^{\text{disagree}}_{c_1,c_2}))\right) \\
&= H_2(0.5) - \left(\nu^{\text{agree}}_{c_1,c_2} H_2(\eta^{\text{agree}}_{c_1,c_2} + \nu^{\text{disagree}}_{c_1,c_2} H_2(\eta^{\text{disagree}}_{c_1,c_2})\right) \\
&= H_2(0.5) - \left(\nu^{\text{agree}}_{c_1,c_2} H_2\left(\frac{c_1 c_2}{\nu^{\text{agree}}_{c_1,c_2}}\right) + \nu^{\text{disagree}}_{c_1,c_2} H_2\left(\frac{c_1(1-c_2)}{\nu^{\text{disagree}}_{c_1,c_2}}\right)\right) \\
&= H_2(0.5) - \left(\nu^{\text{agree}}_{c_1,c_2} H_2\left(\frac{c_1 c_2}{\nu^{\text{agree}}_{c_1,c_2}}\right) + \nu^{\text{disagree}}_{c_1,c_2} H_2\left(\frac{c_1(1-c_2)}{\nu^{\text{disagree}}_{c_1,c_2}}\right)\right) \\
&= H_2(0.5) - \left(\nu^{\text{agree}}_{c_1,c_2}\frac{c_1 c_2}{\nu^{\text{agree}}_{c_1,c_2}} \log\left(\frac{\nu^{\text{agree}}_{c_1,c_2}}{c_1 c_2}\right)\right. \\
&\qquad + \nu^{\text{agree}}_{c_1,c_2}\frac{(1-c_1)(1-c_2)}{\nu^{\text{agree}}_{c_1,c_2}} \log\left(\frac{\nu^{\text{agree}}_{c_1,c_2}}{(1-c_1)(1-c_2)}\right) \\
&\qquad + \nu^{\text{disagree}}_{c_1,c_2}\frac{c_1(1-c_2)}{\nu^{\text{disagree}}_{c_1,c_2}} \log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{c_1(1-c_2)}\right) \\
&\qquad \left. + \nu^{\text{disagree}}_{c_1,c_2}\frac{(1-c_1)c_2}{\nu^{\text{disagree}}_{c_1,c_2}} \log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{(1-c_1)c_2}\right)\right) \\
&= H_2(0.5) - \left(c_1 c_2 \log\left(\frac{\nu^{\text{agree}}_{c_1,c_2}}{c_1 c_2}\right)\right. \\
&\qquad + (1-c_1)(1-c_2) \log\left(\frac{\nu^{\text{agree}}_{c_1,c_2}}{(1-c_1)(1-c_2)}\right) \\
&\qquad + c_1(1-c_2) \log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{c_1(1-c_2)}\right) \\
&\qquad \left. + (1-c_1)c_2 \log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{(1-c_1)c_2}\right)\right)
\end{aligned}$$

The first derivative is

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}c_1}\phi_{c_2}(c_1) = 0 - \Bigg(& c_2 \log\left(\frac{\nu^{\text{agree}}_{c_1,c_2}}{c_1 c_2}\right) - \frac{c_2(1-c_2)}{\nu^{\text{agree}}_{c_1,c_2}} \\
& - (1-c_2)\log\left(\frac{\nu^{\text{agree}}_{c_1,c_2}}{(1-c_1)(1-c_2)}\right) + \frac{c_2(1-c_2)}{\nu^{\text{agree}}_{c_1,c_2}} \\
& + (1-c_2)\log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{c_1(1-c_2)}\right) - \frac{c_2(1-c_2)}{\nu^{\text{disagree}}_{c_1,c_2}} \\
& - c_2 \log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{(1-c_1)c_2}\right) + \frac{c_2(1-c_2)}{\nu^{\text{disagree}}_{c_1,c_2}}\Bigg) \\
= -\Bigg(& c_2 \log\left(\frac{\nu^{\text{agree}}_{c_1,c_2}}{c_1 c_2}\right) - (1-c_2)\log\left(\frac{\nu^{\text{agree}}_{c_1,c_2}}{(1-c_1)(1-c_2)}\right) \\
& + (1-c_2)\log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{c_1(1-c_2)}\right) - c_2 \log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{(1-c_1)c_2}\right)\Bigg) \\
= -\Big(& c_2 \log(\nu^{\text{agree}}_{c_1,c_2}) - c_2\log(c_1) - c_2\log(1-c_2) \\
& - (1-c_2)\log(\nu^{\text{agree}}_{c_1,c_2}) + (1-c_2)\log(1-c_1) + (1-c_2)\log(1-c_2) \\
& + (1-c_2)\log(\nu^{\text{disagree}}_{c_1,c_2}) - (1-c_2)\log(c_1) - (1-c_2)\log(1-c_2) \\
& - c_2\log(\nu^{\text{disagree}}_{c_1,c_2}) + c_2\log(1-c_1) + c_2\log(c_2)\Big) \\
= \log & \left(\frac{c_1}{1-c_1}\right) + (2c_2-1)\log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{\nu^{\text{agree}}_{c_1,c_2}}\right).
\end{aligned}$$

The second derivative is

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}^2 c_1}\phi_{c_2}(c_1) &= \frac{1}{c_1(1-c_1)} + (2c_2-1)\frac{\nu^{\text{agree}}_{c_1,c_2}}{\nu^{\text{disagree}}_{c_1,c_2}} \cdot \frac{(2c_2-1)\nu^{\text{disagree}}_{c_1,c_2} - \nu^{\text{agree}}_{c_1,c_2}(1-2c_2)}{(\nu^{\text{agree}}_{c_1,c_2})^2} \\
&= \frac{1}{c_1(1-c_1)} + (2c_2-1)^2\frac{1}{\nu^{\text{disagree}}_{c_1,c_2}} \cdot \frac{\nu^{\text{disagree}}_{c_1,c_2} + \nu^{\text{agree}}_{c_1,c_2}}{\nu^{\text{agree}}_{c_1,c_2}} \\
&= \frac{1}{c_1(1-c_1)} + \frac{(2c_2-1)^2}{\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}} \\
&= \frac{\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2} + (2c_2-1)^2 c_1(1-c_1)}{c_1(1-c_1)\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}} \\
&= \frac{c_2(1-c_2)}{c_1(1-c_1)\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}}.
\end{aligned}$$

The second derivative is non-negative for $c_2 \in [0.5, 1], c_1 \in [0.5, 1)$. Thus,

$\sum_{c_2 \in \Omega_{f_2}} f_2(c_2)\phi_{c_2}(c_1)$ is a convex scoring function. With Corollary S7 follows the desired statement. □

In the last proposition, we assume that only the individual classifier's information is constraint but not their accuracy and look at the resulting ensemble information.

Proposition 9. ***(Information constrained specialists and generalists bound the ensemble information)*** *Consider k classifiers with individual information I_i and confidence distributions f_i ($i \in \{1..k\}$). For each i, let the accuracies corresponding to the individual information be $\tilde{acc}_i = H_2^{-1}(1 - I_i)$. Let $f^{generalist}_{\underset{\sim}{acc}_i}$ and $f^{specialist}_{\underset{\sim}{acc}_i}$ as defined above. Now consider the ensemble information based on the original classifiers, with ensemble confidence distribution $f_e = \bigotimes_{i=1}^k f_i$ as well as ensembles with confidence distributions $\tilde{f}_e^{generalist} = \bigotimes_{i=1}^k f^{generalist}_{\underset{\sim}{acc}_i}$ and $\tilde{f}_e^{specialist} = \bigotimes_{i=1}^k f^{specialist}_{\underset{\sim}{acc}_i}$ as in Theorem 7. The information of the ensemble classifier is bounded by*

$$I_{\tilde{f}_e^{generalist}} \leq I_{f_e} \leq I_{\tilde{f}_e^{specialist}}.$$

Proof. The ensemble information is $\mathrm{I}_f = I(Y; O_e)$ as in Proposition 8. Our main work in this proof is to show that the function in the double sum is convex in $1 - H_2(c_1)$ so that we can apply Corollary S7 again. (In Proposition 8 we have only shown that it is convex in c_1.) Also denote the local information by $\iota(c) = H_2(0.5) - H_2(c)$. The relevant term in the big brackets is

$$\phi_{c_2}(c_1) = \left(\nu^{\text{agree}}_{c_1,c_2}\iota(\eta^{\text{agree}}_{c_1,c_2}) + \nu^{\text{disagree}}_{c_1,c_2}\iota(\eta^{\text{disagree}}_{c_1,c_2})\right)$$

We will show that $\phi_{c_2}(c_1)$ is convex in $\iota(c_1)$ by showing that the second derivative is non-negative.

$$\begin{aligned}
\frac{\mathrm{d}^2\phi_{c_2}(c_1)}{\mathrm{d}\iota(c_1)^2} &= \frac{\mathrm{d}}{\mathrm{d}\iota(c_1)}\frac{\mathrm{d}\phi_{c_2}(c_1)}{\mathrm{d}\iota(c_1)} \\
&= \frac{\mathrm{d}}{\mathrm{d}\iota(c_1)}\left(\frac{\frac{\mathrm{d}\phi_{c_2}(c_1)}{\mathrm{d}c_1}}{\frac{\mathrm{d}\iota(c_1)}{\mathrm{d}c_1}}\right) \\
&= \frac{\phi''_{c_2}\iota' - \phi'_{c_2}\iota''}{(\iota')^3}
\end{aligned}$$

We had already derived ι', ι'', $\phi'_{c_2}c_1$ and $\phi''_{c_2}c_1$ in the proof of Proposition 8:

$$\begin{aligned}
\iota'(c) &= \frac{\mathrm{d}}{\mathrm{d}c}\iota(c) = \log\left(\frac{c}{1-c}\right) \\
\iota''(c) &= \frac{\mathrm{d}^2}{\mathrm{d}c^2}\iota(c) = \frac{1}{c(1-c)} \\
\phi'_{c_2}c_1 &= \frac{\mathrm{d}}{\mathrm{d}c_1}\phi_{c_2}(c_1) = \log\left(\frac{c_1}{1-c_1}\right) + (2c_2-1)\log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{\nu^{\text{agree}}_{c_1,c_2}}\right) \\
\phi''_{c_2}c_1 &= \frac{\mathrm{d}^2}{\mathrm{d}^2c_1}\phi_{c_2}(c_1) = \frac{c_2(1-c_2)}{c_1(1-c_1)\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}}.
\end{aligned}$$

Now we can put together the second derivative of $\phi_{c_2}(c_1)$ with respect to $\iota(c_1)$. For convexity, we want to show that this is non-negative.

$$\begin{aligned}
&\frac{\mathrm{d}^2\phi_{c_2}(c_1)}{\mathrm{d}\iota(c_1)^2} \geq 0 \\
&\iff \frac{\phi''_{c_2}\iota' - \phi'_{c_2}\iota''}{(\iota')^3} \geq 0 \\
&\overset{(1)}{\iff} \phi''_{c_2}\iota' - \phi'_{c_2}\iota'' \geq 0 \\
&\iff \frac{c_2(1-c_2)}{c_1(1-c_1)\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}}\log\left(\frac{c_1}{1-c_1}\right) \\
&\qquad - \left(\log\left(\frac{c_1}{1-c_1}\right) + (2c_2-1)\log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{\nu^{\text{agree}}_{c_1,c_2}}\right)\right)\frac{1}{c_1(1-c_1)} \geq 0 \\
&\overset{(2)}{\iff} \frac{c_2(1-c_2)}{\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}}\log\left(\frac{c_1}{1-c_1}\right) \\
&\qquad - \left(\log\left(\frac{c_1}{1-c_1}\right) + (2c_2-1)\log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{\nu^{\text{agree}}_{c_1,c_2}}\right)\right) \geq 0 \\
&\iff \log\left(\frac{c_1}{1-c_1}\right)\left(\frac{c_2(1-c_2)}{\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}} - 1\right) - (2c_2-1)\log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{\nu^{\text{agree}}_{c_1,c_2}}\right) \geq 0 \\
&\iff \log\left(\frac{c_1}{1-c_1}\right)\left(\frac{c_2(1-c_2) - \nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}}{\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}}\right) \\
&\qquad - (2c_2-1)\log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{\nu^{\text{agree}}_{c_1,c_2}}\right) \geq 0
\end{aligned}$$

$$\iff \log\left(\frac{c_1}{1-c_1}\right)\left(\frac{(2c_2-1)^2c_1(c_1-1)}{\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}}\right) - (2c_2-1)\log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{\nu^{\text{agree}}_{c_1,c_2}}\right) \geq 0$$

$$\overset{(3)}{\iff} \log\left(\frac{c_1}{1-c_1}\right)\left(\frac{(2c_2-1)c_1(c_1-1)}{\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}}\right) - \log\left(\frac{\nu^{\text{disagree}}_{c_1,c_2}}{\nu^{\text{agree}}_{c_1,c_2}}\right) \geq 0$$

$$\iff \log\left(\frac{\nu^{\text{agree}}_{c_1,c_2}}{\nu^{\text{disagree}}_{c_1,c_2}}\right) - \log\left(\frac{c_1}{1-c_1}\right)\left(\frac{(2c_2-1)c_1(1-c_1)}{\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}}\right) \geq 0$$

In (1), (2) and (3) we multiply with $(\iota')^3$, $\frac{1}{c_1(1-c_1)}$ and $\frac{1}{2c_2-1}$, respectively. These terms are larger than 0 in $c_1, c_2 \in (0.5, 1)$ so that the inequality sign does not change. It remains to show that

$$\omega(c_1,c_2) := \log\left(\frac{\nu^{\text{agree}}_{c_1,c_2}}{\nu^{\text{disagree}}_{c_1,c_2}}\right) - \log\left(\frac{c_1}{1-c_1}\right)\left(\frac{(2c_2-1)c_1(1-c_1)}{\nu^{\text{disagree}}_{c_1,c_2}\nu^{\text{agree}}_{c_1,c_2}}\right) \geq 0.$$

We will do so by switching to the partial derivative with respect to c_2 (instead of c_1 as above) and demonstrate that (i) $\omega(c_1,c_2) = 0$ for $c_2 = 0.5$ and $c_2 = 1$, (ii) $\frac{\partial\omega(c_1,c_2)}{\partial c_2}\Big|_{c_2=0.5} \geq 0$, and (iii) $\frac{\partial\omega(c_1,c_2)}{\partial c_2} = 0$ for only one $c_2 \in (0.5, 1)$.

(i) We show that ω is 0 at the edge cases.

$$\begin{aligned}
\omega(c_1, 0.5) &= \log\left(\frac{0.5}{0.5}\right) - \log\left(\frac{c_1}{1-c_1}\right)\left(\frac{(2\cdot 0.5-1)c_1(1-c_1)}{0.5\cdot 0.5}\right) \\
&= 0 - 0 = 0 \\
\omega(c_1, 1) &= \log\left(\frac{c_1}{1-c_1}\right) - \log\left(\frac{c_1}{1-c_1}\right)\left(\frac{(2\cdot 1-1)c_1(1-c_1)}{c_1(1-c_1)}\right) \\
&= \log\left(\frac{c_1}{1-c_1}\right) - \log\left(\frac{c_1}{1-c_1}\right) = 0
\end{aligned}$$

(ii) We show that ω starts with a nonnegative slope.

$$\begin{aligned}
\frac{\partial\omega(c_1,c_2)}{\partial c_2} &= \frac{\nu^{\text{disagree}}_{c_1,c_2}}{\nu^{\text{agree}}_{c_1,c_2}} \cdot \frac{(2c_1-1)\nu^{\text{disagree}}_{c_1,c_2} - \nu^{\text{agree}}_{c_1,c_2}(-1)(2c_1-1)}{(\nu^{\text{disagree}}_{c_1,c_2})^2} \\
&\quad - \log\left(\frac{c_1}{1-c_1}\right) c_1(1-c_1) \\
&\quad \cdot \left(\frac{2\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2} - (2c_2-1)(-1)(2c_2-1)(2c_1-1)^2}{(\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2})^2}\right) \\
&= \frac{2c_1-1}{\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2}} - \log\left(\frac{c_1}{1-c_1}\right) c_1(1-c_1) \\
&\quad \cdot \left(\frac{2\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2} + (2c_2-1)^2(2c_1-1)^2}{(\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2})^2}\right) \\
\frac{\partial\omega(c_1,c_2)}{\partial c_2}\Big|_{c_2=0.5} &= \frac{2c_1-1}{0.5\cdot 0.5} - \log\left(\frac{c_1}{1-c_1}\right) c_1(1-c_1) \\
&\quad \cdot \left(\frac{2\cdot 0.5\cdot 0.5 + (2\cdot 0.5-1)^2(2c_1-1)^2}{(0.5\cdot 0.5)^2}\right) \\
&= (8c_1-4) - 8c_1(1-c_1)\log\left(\frac{c_1}{1-c_1}\right)
\end{aligned}$$

To finish (ii) we have to show that $\widetilde{\omega}(c_1) := \frac{\partial\omega(c_1,c_2)}{\partial c_2}\Big|_{c_2=0.5} \geq 0$. We use a similar strategy as before: We show that (ii.i) $\widetilde{\omega}(0.5) = 0$ and that (ii.ii) $\frac{\mathrm{d}\widetilde{\omega}(c_1)}{\mathrm{d}c_1} \geq 0$. This way, we are using the same proof strategy again nested within the first.

(ii.i) We show that $\widetilde{\omega}$ starts at 0.

$$\widetilde{\omega}(0.5) = (8\cdot 0.5-4) - 8\cdot 0.5\cdot(1-0.5)\log\left(\frac{0.5}{1-0.5}\right) = 0 - 2\cdot 0 = 0$$

(ii.ii) We show that $\widetilde{\omega}$ has a nonnegative slope.

$$\begin{aligned}
\frac{\mathrm{d}\widetilde{\omega}(c_1)}{\mathrm{d}c_1} &= 8 - 8\left((1-2c_1)\log\left(\frac{c_1}{1-c_1}\right)\right. \\
&\quad \left. + c_1(1-c_1)\frac{1-c_1}{c_1}\frac{1\cdot(1-c_1)-(-c_1)}{(1-c_1)^2}\right) \\
&= 8 - 8\left((1-2c_1)\log\left(\frac{c_1}{1-c_1}\right) + 1\right) \\
&= 8(2c_1-1)\log\left(\frac{c_1}{1-c_1}\right) \geq 0
\end{aligned}$$

Taken together, $\widetilde{\omega}(c_1)$ starts non-negative at $c_1 = 0.5$ (ii.i) and only increases for larger c_1 (ii.ii). This finishes (ii) showing that

$$\widetilde{\omega}(c_1) = \frac{\partial \omega(c_1, c_2)}{\partial c_2}\Big|_{c_2=0.5} = (8c_1 - 4) - 8c_1(1-c_1)\log\left(\frac{c_1}{1-c_1}\right) \geq 0$$

(iii) It remains to show that ω changes monotonicity only once. We do so by finding the unique zero for the first derivative.

$$\begin{aligned}
&\frac{\partial \omega(c_1, c_2)}{\partial c_2} \stackrel{!}{=} 0 \\
\iff &\frac{2c_1 - 1}{\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2}} \\
&\quad - \log\left(\frac{c_1}{1-c_1}\right) c_1(1-c_1)\left(\frac{2\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2} + (2c_2-1)^2(2c_1-1)^2}{(\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2})^2}\right) \stackrel{!}{=} 0 \\
\stackrel{(1)}{\iff} &(2c_1-1)\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2} \\
&\quad - \log\left(\frac{c_1}{1-c_1}\right) c_1(1-c_1)\left(2\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2} + (2c_2-1)^2(2c_1-1)^2\right) \stackrel{!}{=} 0 \\
\iff &\left((-8c_1^3 + 12c_1^2 - 6c_1 + 1) - c_1(1-c_1)\log\left(\frac{c_1}{1-c_1}\right)(8c_1^2 - 8c_1 + 2)\right)c_2^2 \\
&\quad - \left((-8c_1^3 + 12c_1^2 - 6c_1 + 1) - c_1(1-c_1)\log\left(\frac{c_1}{1-c_1}\right)(8c_1^2 - 8c_1 + 2)\right)c_2 \\
&\quad + \left((-2c_1^3 + 3c_1^2 - c_1) - c_1(1-c_1)\log\left(\frac{c_1}{1-c_1}\right)(2c_1^2 - 2c_1 + 1)\right) \stackrel{!}{=} 0 \\
\stackrel{(2)}{\iff} &c_2^2 - c_2 \\
&\quad + \frac{\left((-2c_1^3 + 3c_1^2 - c_1) - c_1(1-c_1)\log\left(\frac{c_1}{1-c_1}\right)(2c_1^2 - 2c_1 + 1)\right)}{\left((-8c_1^3 + 12c_1^2 - 6c_1 + 1) - c_1(1-c_1)\log\left(\frac{c_1}{1-c_1}\right)(8c_1^2 - 8c_1 + 2)\right)} \stackrel{!}{=} 0
\end{aligned}$$

At (1) we multiply with $(\nu^{\text{agree}}_{c_1,c_2}\nu^{\text{disagree}}_{c_1,c_2})^2$ and in (2) we divide by the constant in the denominator. Both are larger than 0 in $c_1, c_2 \in (0.5, 1)$. The result is a quadratic equation that has at most one zero in $c_2 \in (0.5, 1)$. This completes (iii).

Taken together, for any $c_1 \in (0.5, 1)$, $\omega(c_1, c_2)$ is zero at the corner cases $c_2 = 0.5$ and $c_2 = 1$ (i), increases from $c_2 = 0.5$ on (ii) and only changes monotonicity once (iii) so that $\omega(c_1, c_2)$ is non-negative.

With that, ϕ_{c_2} is convex in $\iota(c_1)$. Lemma S6 yields the desired result: The ensemble information is convex in the individual information so that specialists maximize and generalists minimize the ensemble information.

□

4.7 Discussion

In a setting in which individual classifiers output predictions together with confidences (probability elicitation), we have shown that the accuracy of the ensemble depends on the exact confidence distributions. Classifiers that distinguish between high and low confidences perform better than those that always produce moderate confidences. We have provided bounds when (a) only the individual classifiers' accuracies are known and (b) when both, the individual classifiers' accuracies and mutual information, are known. These bounds can be used to determine how many classifiers must be included in an ensemble to guarantee a target accuracy, see Figure 4.5.

For our running example, this means that even if we know how often a classifier can predict correctly whether a patient has a disease or not based on a single retina image, we cannot uniquely determine the accuracy of an ensemble of such classifiers. However, we can provide bounds and improve on these bounds when we additionally know the transmitted information of these classifiers.

Classifiers in an ensemble should ideally be constructed such that they specialize: For a given accuracy, ideal classifiers should sometimes predict with close to absolute certainty even if this comes at the cost of not learning on other parts of the input space. The advantage of specialists comes through despite the specialists not coordinating on which areas of the input space they specialize.

Chapter 5

Discussion

We have analyzed methods in two fields of psychological research through the lens of Information Theory. In unconscious priming, we have revealed a flaw in a standard reasoning that produced many unwarranted claims about unconscious processing. Our more appropriate method, the sensitivity comparison, shows that there is often no evidence for processing that goes beyond what participants can consciously report.

In group decision making, real group decisions are often found to be superior to simulating group decisions generated from statistically aggregating individual reports. We took the theoretically optimal method, Confidence Weighted Majority Voting (CWMV), and showed in an experiment that simulations with this method perform as well as real group decisions. Differences between real vs. simulated group decisions may often stem from methodological issues and vanish when using optimal methods that incorporate all available information. In both fields, individual human participants' responses carry information that must be taken into account when making comparisons with either indirect measures in priming or—after statistically aggregating them into simulated group decisions—with real group decisions.

Based on these observations, we asked whether the accuracy of decisions made by groups can be determined by the accuracies of their individual members. We found that individual responses can carry vastly different amounts of information even if their accuracy is held constant. Consequently, the accuracy of a group can take values in a wide range even if the accuracy of the individuals is known. We have quantified this range by proving bounds for the best- and worst-case group (ensemble) accuracy and discussed guiding principles for the construction or selection of individual classifiers to maximize group performance.

5.1 Renewed Skepticism in Unconsciousness Research

In the first field, unconscious priming, we found that previous studies routinely applied a standard reasoning that inappropriately compared measures of conscious and unconscious information processing. The problem with the standard reasoning is that this comparison is biased towards better unconscious than conscious information processing for purely methodological reasons. This has led to the wrong impression that human participants process more information than they can report consciously. In other words, it was erroneously concluded that processing occurred outside of conscious awareness. We developed a method to reanalyze those studies. In this method, the sensitivity comparison, the information considered in both measures is equated and properly compared such that no misleading differences appear. Our appropriate method reveals that the empirical basis for claims about unconscious processing in many cases is missing. Unconscious information processing does not seem as ubiquitous as previously thought. On the contrary, many studies that claimed evidence for unconscious processing seem to, on closer look, provide no tangible evidence for their claims. Interpretations about unconscious processing have to be reevaluated by the field, which has a substantial impact on our theoretical reasoning about conscious vs. unconscious processing.

For our reanalysis, we had to make one critical assumption about an underlying variance ratio, q^2. This was necessary to estimate the accuracy from the indirect (unconscious) measures. We were reasonably conservative with this assumption and considered multiple sources in order to estimate it (Section 2.6.4). Nevertheless, it is possible that our conservative assumption may be violated and the true q^2 may be larger in some situations. In these cases, our reanalysis may not confirm previous claims about a difference between conscious vs. unconscious measures even though there are true differences. To solve this problem, the original full trial-by-trial data is required to perform our sensitivity comparison and to determine without additional assumptions whether there is a difference between the two tasks.

Future research must use our or similar methods to avoid the fallacies pointed out here and to corroborate claims about unconscious processing. For the stimuli and setups of the studies that we reanalyzed, the question of whether processing occurred unconsciously is up in the air again.

One fundamental open question is whether the operational definition of consciousness should be subjective or objective (Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008): Does unconscious processing occur when participants have the impression that they did not process anything (subjective)

or when their discrimination performance is at chance level guessing (objective)? In many studies using the priming paradigm, participants would claim that they were not subjectively aware of the stimuli even though their guesses about the category of the stimuli were correct above chance level. This is not a contradiction but shows that the two definitions are not equal: Subjectively, participants report not seeing stimuli in some situations even if they can still discriminate them above chance level when asked to guess (King & Dehaene, 2014; Stein, Kaiser, Fahrenfort, & Van Gaal, 2021). Thus, subjective measures are less restrictive and may produce different results: Information processing may be subjectively unconscious even if participants have full introspective access to that processing, which may simply be too weak to produce a noteworthy subjective experience. Researchers do not agree on what the one true measure for consciousness is, let alone on its definition (Irvine, 2013, 2017; Rothkirch & Hesselmann, 2017). We do not take a position in this discussion. In the foreseeable future, consciousness remains a fractured concept without a unified definition.

However, most studies that use the standard reasoning commit to objective measures. They measure conscious awareness through discrimination performance in the direct task thereby using objective measures of consciousness in form of accuracies (%-correct) or sensitivities (d'). We simply point out that, when objective measures are used, they should be evaluated in a consistent way. When comparing accuracies or sensitivities from the direct task with measures from the indirect task, the appropriate method is to obtain the same metrics obtained from the indirect measures and tests for a difference.

It has been previously noted that objective measures of awareness are not measured with sufficient precision (Vadillo et al., 2020; Buchner & Wippich, 2000). Our point goes beyond this critique. Our sensitivity comparison requires not only that the objective awareness measures are sufficiently precise but also how to evaluate them: by a comparison between direct vs. indirect tasks. Only a difference can be the basis for further interpretation about unconsciousness (in the sense of a necessary but not sufficient condition). Our method follows recent demands to improve general methodology by measuring effect sizes, whether it be in the framework of Frequentists (Cumming, 2014) or Bayesians (Kruschke, 2013; Dienes, 2014; Kruschke & Liddell, 2018). Simply evaluating both tasks in isolation is no solution, neither for Frequentists (Gelman & Stern, 2006) nor for Bayesians (Palfi & Dienes, 2020).

The critique we develop for the unconscious priming paradigm extends to other paradigms. Since priming has been one of the first methods to empirically demonstrate unconscious processing (Eriksen, 1960), many followed in its wake (Kim & Blake, 2005, see their Table 1 on p. 385 for an instructive overview).

Many of the new paradigms are affected by similar problems.

For example, consider the *contextual cueing* paradigm (Chun, 2000; Colagiuri & Livesey, 2016). In a typical experiment, participants see a configuration of multiple stimuli. Their task is to find one odd stimulus among the others, for example, to find a T among a display of Ls. This is done in multiple hundreds of trials. Some of the configurations are repeated across trials so that the T stimulus is always at the same location. Other configurations are randomly constructed so that participants have not seen them before. Over the course of the experiment, participants show increasingly faster reaction times in the repeated configurations as compared with the new configurations. This is evidence that participants, on some level, recognized the repeated configurations. This reaction time effect between repeated vs. new configurations is comparable to the reaction time effect between congruent vs. incongruent stimuli from priming studies. After the experiment proper, participants are presented with multiple configurations and have to discriminate which of them they have seen before and which are new. In this task, participants perform close to chance—just as in the direct task in the unconscious priming paradigm. From this result, it is often concluded that there was some recognition of the repeated configurations but that this recognition was unconscious because participants cannot reliably discriminate which ones they recognized.

The contextual cueing paradigm produces the same pattern of results as unconscious priming, makes the same interpretation and, thus, suffers from the same fundamental flaw. The typical pattern of results seems to indicate processing beyond what participants can report when directly asked. But this comparison is improper. The reaction time effects found in the visual search task may be entirely explained by a weak, residual conscious recognition of the repeated configurations. Our critique from Chapter 2 can be transferred directly to these studies (e.g., to the seminal study by Chun & Jiang, 1998 but also to more recent studies such as Schlagbauer, Rausch, Zehetleitner, Müller, & Geyer, 2018; Smyth & Shanks, 2008; Zhao & Ren, 2020). To find genuine evidence for unconscious information processing (implicit recognition) one has to compare objective measures between the two tasks and establish an empirical difference. For example, if one would find that reaction times are suited to determine which configurations were repeated vs. new with a higher accuracy than participants' direct responses, then there is an empirical basis from which (with additional theoretical assumptions) evidence for unconscious processing can be derived. Future research has to implement such methods. Even when arguing with correlations as Colagiuri and Livesey (2016) do in this paradigm, arguments of the dissociation paradigm (Schmidt & Vorberg, 2006) carry over and still require our analysis to investigate the occurrence of

superior indirect measures.

In other paradigms, critical perspectives on unconscious processing have emerged as well. Here are a few examples in short: In continuous flash suppression, prominent studies have reported that human participants are able to read and solve equations unconsciously (Sklar et al., 2012) or discriminate objects from backgrounds (Mudrik, Breska, Lamy, & Deouell, 2011). But this was soon heavily criticized by Moors and Hesselmann (2018); Hesselmann and Moors (2015) and citeAmoors2016scene, respectively. In memory research, Buchner and Wippich (2000) criticized that conscious awareness was not measured precisely enough (similar to the critique from Vadillo et al., 2020 in unconscious learning). In (individual) decision making, the prominent *unconscious thought theory* (UTT; Dijksterhuis et al., 2006) has emerged positing that deliberation without conscious thoughts can outperform conscious deliberations. But Newell and Shanks (2014) introduce serious doubts in the results of this line of research as well. Taken together, many strands of research seem to have produced evidence for unconscious processing—perhaps driven by enthusiasm about the power of unconscious thought. Upon scrutiny, each of these results appears less convincing leaving surprisingly little tangible evidence for unconscious processing.

In the century-long debate about conscious vs. unconscious processes in psychology, it seems that "the pendulum is now slowly swinging back to a state where one can again doubt that complex information processing can take place unconsciously" (Cleeremans & Tallon-Baudry, 2021, p. 2). This skeptical perspective has last been prominently held by Merikle and Reingold (1998); Merikle (1992). After that and during the last twenty years, the general notion swung in favor of unconscious processing. This can be partly explained by the rise of neural imaging measures that can indirectly assess participants' information processing, namely, via electroencephalography (EEG), functional magnetic resonance imaging (fMRI), etc. Enthusiasm sparked by results from these measures prompted the notion that, when in doubt, one should assume that some process can occur outside of conscious awareness (Hassin, 2013). But this notion is now again replaced a more skeptical perspective (Phillips, 2021b) in which it is held that participants have rich introspective access to the information they processed (Peters & Lau, 2015; Phillips, 2021a).

5.2 Group Decision Making Requires Unified Models of Dependencies

In the second field, group decision making, we observed a similar problem as in unconscious priming. In this field, responses from individual participants are statistically aggregated into a simulated group decision. These simulated group decisions are meant to substitute real group decisions but simulated group performance often falls short of real group performance. One reason for this is that, again, some information from the individual participants is discarded. Often, simulations do not take into account confidences of the individual reports in the optimal way. In these cases, it should come as no surprise that simulated group decisions are inferior to real group decisions because, by design, less information is incorporated in the simulations. We put forward the theoretically optimal method for these simulations, Confidence Weighted Majority Voting (CWMV). We investigated this method in an experiment showing that CWMV simulations can perform on par with real groups. However, real groups deviated from simulations with this method. We adapted the simulation method to better match real groups in order to gain a better understanding of the inner workings of real group discussions. Our adaptation works on the level of log odds which is often appropriate to model decisions under uncertainty (Zhang & Maloney, 2012).

The main limitation of our experiment is its simplified setting. We used stimuli of a simple and stochastic nature to manipulate individual confidence reports and make them *independent*. In reality, confidences arise in much less controlled settings. Confidences from individuals will have dependencies for example because individuals formed their decisions based on shared material. This poses a problem to CWMV because this method fails to simulate group decisions as violations of the independence assumption occur.

The independence assumptions on individual confidences are not only built into CWMV (see Nitzan & Paroush, 1984 for discussion of this assumption) but are inextricably ingrained in virtually all statistical and machine learning applications. It usually comes in the form of the i.i.d. (independently and identically distributed) assumption (Herzog, Francis, & Clarke, 2019; Pishro-Nik, 2014; Murphy, 2012; Winer et al., 1991, to name a few established text books). The difficulty with relaxing the independence assumption is that there is not only one kind of dependence. Rejecting independence, one has to specify which type of dependence exists in the data. This then produces methods tailored to specific kinds of dependencies that do not generalize to other kinds.

For example, Kaniovski and Zaigraev (2011) relax the independence assumption in CWMV using a model with correlated votes (see also Ladha, 1992

and see Bahadur, 1961 for the theorem that makes these approaches computationally appealing). Correlations are a widely used measure for dependence and non-linearity concerns are not applicable in binary decision cases. However, pairwise correlations do not capture dependencies exhaustively in groups with multiple individuals: There are situations in which all pairwise correlations are 0 but higher-order correlations reveal a complete dependence of one of the members' decision patterns. Thus, to fully capture the dependencies with this model an exponential number of parameters has to be fitted (all correlations of the third order, fourth order, etc.). Yet another alternative from Shapley and Grofman (1984) is to assume a model in which a fixed number of randomly chosen individuals "blackout" giving an incorrect decision in these cases. Finally, Estlund (1994) suggests a model in which individuals tend to follow an opinion leader. All these models are rather specific and may not generalize well to other contexts.

One approach that appears particularly appealing to us are *hidden profiles* (introduced in Stasser & Titus, 2003, 1985; see Lu, Yuan, & McLeod, 2012 for a review). In this paradigm, shared vs. unique information that is held commonly among group members vs. uniquely by only one group member is manipulated. The deliberate distribution of shared vs. unique information introduces dependencies between individuals' reports. Thus, the method of CWMV would fail when applied to the individuals reports to simulate a group decision. The solution we discussed in Chapter 3 does not adapt the method but what the method is applied to: Instead of weighting individuals' decisions which are not dependent, CWMV should be applied to the pieces of information that are held by the individuals. This way, shared information does not enter the simulated group decision as many times as there are individuals who know it but only once. This defers the problem from adapting the method to finding a suitable level of abstraction has to be found on which independence assumptions can be made (e.g., individual pieces of information instead of individual group members). In how far CWMV can make accurate simulations for these different abstraction levels remains an open research question.

To conclude this part, we want to emphasize that we avoided normative language when presenting results in Chapter 3: Instead of using the term *equality bias* (as in Mahmoodi et al., 2015) we refer to our result as *equality effect*. Even though CWMV is the theoretically optimal method, human group discussions do not exist in a theoretical vacuum. Instead, discussions are influenced by other factors such as social aspects that may explain why groups tend to give individuals a more equal say in the real group decision as compared to the CWMV simulations. Austen-Smith and Banks (1996) argued that it is not necessarily the case that group members make the same reports individually

as they do in the real group discussion; it may be rational for an individual to present different information in order to pivot the group decision towards their beliefs. Such motives are also not captured in our simplified experiment. We, therefore, remain cautious to make further interpretations. Our main goal was not to determine in which ways groups behave suboptimally but to place emphasis on the CWMV method for simulating group decisions.

5.3 Finding Applications for Ensemble Accuracy Bounds

From the work on the two psychological fields, we arrived at a more theoretical research question: Can we predict the accuracy of a group when we know the accuracies of the individual members of that group? This is relevant not only in psychology but also in machine learning, where classifiers are often combined into a group, or ensemble, to increase accuracy—be it in random forests (Breiman, 2001) or other ensemble methods (Hansen & Salamon, 1990; Lakshminarayanan et al., 2017). We have demonstrated that the ensemble accuracy is elusive in the sense that the accuracies of the individuals do not uniquely determine the accuracy of the ensemble. In contrast, the ensemble accuracy can take values in a surprisingly wide range.

One major aspect underlying differences in transmitted information among individuals with the same accuracy are their confidence distributions. Individuals provide more information when they can give differential confidences (sometimes high vs. sometimes low) rather than undifferentiated confidences (always moderate). This meta-information then affects the ensemble accuracy. Thus, individuals' information—the information between individuals' responses and the to be predicted label—is an indicator of good ensemble accuracy. Consequently, we show that when individuals' information are additionally known, we can narrow down the range of possible ensemble accuracies.

Our bounds on the performance of an ensemble given the performance of the individuals have two immediate applications. First, they can determine what we should expect from ensemble classifiers but also from real group discussions. Take as an example the study from Klein and Epley (2015) in which we can predict the range of group accuracies with our bounds. Klein & Epley compared real groups to statistical aggregations of individual responses in a lie-detection task. Individuals saw videos of suspects that were lying or telling the truth. They then either gave individual guesses that were statistically combined in a majority vote to a simulated group decision or discussed in a real group producing a real group decision. Individual guesses had an accuracy of

53.6% (Experiment 1). While simulated group decisions from three individuals were only slightly more reliable with an accuracy of 54.5%, real groups reached an accuracy of 61.7%. Our ensemble bounds predict exactly this range. When combining three individuals with accuracy of 53.6%, the worse case (all generalists) group accuracy is 55.4% while the best case (all specialists) group accuracy is 60.0%. This suggests that the difference between simulated and real groups in Klein & Epley's experiment did not come from some intangible synergy but simply from participants' leveraging their confidences in the real group discussions whereas simulations based on MV necessarily produced the worst case. Note that Klein & Epley reported that confidence ratings were not correlated with accuracies justifying their choice of a majority vote. However, methodological concerns discussed in Chapter 3 may point to an unreliable measurement of confidences in their study. Thus, it is nevertheless plausible that participants used confidences in the real group discussions.

Second, our bounds allow determining how many individual classifiers are needed to guarantee a desired ensemble performance. We show that, in order to reach a certain ensemble accuracy, of for example 95%, fewer individual classifiers are needed when their individual accuracy is high rather than low. Furthermore, even when those individual classifiers have the same accuracy, fewer of them are needed when their individual information is high rather than low.

Moreover, our results provide guiding principles for constructing or selecting classifiers by showing how high- vs. low-information classifiers are constituted. High-information classifiers are specialized in the sense that they provide differential confidences: They provide high confidences in situations in which they are specialized and low confidences in the rest. In contrast, less informative classifiers are generalists always predicting with a moderate degree of confidence. An open question remains about how to train machine learning classifiers to be specialists rather than generalists. Selecting features instead of classifiers (to which our model can be easily generalized) is already an established principle (Battiti, 1994; Vergara & Estévez, 2014).

One aspect of our construction principles, specialists vs. generalists, is noteworthy. We have seen that specialists rather than generalists transmit more information given that they have the same accuracy. However, this is not the sole reason that makes specialists superior to generalists in ensembles. There is some inherent advantage of specialists that goes beyond the fact that they transmit more information than generalists. To see this, consider a specialist and a generalist that are matched in their information to, say, 0.5 bit. As a consequence, the specialist has a lower accuracy of only 75% while the generalist has an accuracy of 89%. The difference is due to the generalist

having to compensate the surplus of meta-information from the specialist by an overall higher accuracy. Even though both classifiers now have the same individual information, ensembles of such specialists still produce higher information than ensembles of generalists. The reason for this is that they, by their construction, convey more unique information (Griffith & Koch, 2014; Olbrich, Bertschinger, & Rauh, 2015). As a consequence, ensembles of specialists accumulate information faster propelling an increase in ensemble accuracy. Thus, with sufficient size, ensembles of specialists still produce higher accuracies than ensembles of generalists: Even though individually, the specialists only have an accuracy of 75% vs. generalists with 89%, an ensemble of six of these specialists has an accuracy of 99% and outperforms an ensemble of six generalists who only reach 98%. This demonstrates that specialists have desirable properties beyond the addition of meta-information.

These results are tailored to machine learning ensembles consisting of artificial classifiers. But they are also interesting for groups with humans. If we measure the individual performances first, we can then compare the observed group performance to our bounds and diagnose their group interactions: (1) Real groups may fall short of our expectations. Their performance may be below the lower bound indicating inefficient group interactions. (2) If real groups meet our expectations, expensive and time-consuming group discussions can potentially be replaced with statistical aggregations of the individual reports, see Chapter 3. (3) A real group may surpass our expectation by performing above the upper bound. In these cases, further research has the opportunity to uncover yet unformalized mechanisms underlying such groups.

We made quite strong assumptions to obtain our bounds. The role of these assumptions is important to understand. We assumed perfect calibration of confidences. In most applications, confidences of classifiers are only estimates and do not perfectly reflect the probability of a prediction to be correct. However, we made this assumption to emphasize the relevance of our theoretical results. Our results do not stem from an effect of robustness against estimation errors. They stem from an inherent ambiguity: The accuracy of individual classifiers does not uniquely determine the ensemble performance *even if* the individual confidences are perfectly calibrated. Thus, our results are not to be understood in the sense that specialists are more robust to estimation errors but that their construction principle is inherently beneficial for ensembles.

The second major assumption in our setting is that of independence between the classifier's confidence distributions, see also discussion in Section 5.2. Acknowledging the role of this assumption is crucial to understand why our results are novel. Without this assumption, it is not surprising that specialists outperform generalists in ensembles. When specialists divide the input space

among them so that each input point can be perfectly identified by one specialist, it is easy to see that they outperform generalists. But such coordination would lead to dependent confidence distributions: When one classifier is a specialist for a particular input point, it will be highly confident and conversely others will have a low confidence—because they would have chosen different specializations.

Consider the following setting to exemplify our independence assumption. Assume biology students have to take an exam. The exam is on many different topics, too many for one student to learn them all exhaustively. It will consist of two-alternative single-choice questions. It is a group exam and will be held in groups of three. Each student is graded according to how many questions the group gets right. The strategy for each student is to maximize the group accuracy: But should they learn all topics a little bit (generalist) or focus on just some topics while neglecting others (specialist)? Obviously, if the group assignments were known beforehand, students would do well by distributing the topics among them so that, for each exam question, one student will be the specialist. But this produces dependencies that are at odds with our assumptions. Instead, in line with our independence assumption, groups will be assigned randomly only on the day of the exam. Now it is not so obvious anymore that students should learn as specialists because they may be grouped up in a way that specializations overlap. Our results show that it still is beneficial for students to learn as specialists rather than generalists, which produces substantially better group accuracies.

This independence assumption diverges from the typical machine learning context where classifiers are often trained on different subparts of the input space in order to create (negatively) dependent confidence distributions, which improves ensemble accuracies. This type of dependencies falls under the label ensemble diversity (Kuncheva & Whitaker, 2003) or negative correlation learning (Chen, Cohn, & Yao, 2012; Liu & Yao, 1997, 1999).

Instead, our independence assumption is inspired and therefore aligned with the typical experimental settings in psychophysics (see Chapter 2 and 3). For example, Bahrami et al. (2010) conducted the following experiment: They let multiple participants view barely visible stimuli and let them guess as a group the identity of these stimuli. In their setting, the individual participants see the same physical stimuli but still produce varying degrees of confidence. Plausibly, these confidences show no dependence because the noise is entirely located within the visual system of the participants: Higher and lower confidences are only determined by fluctuations of their individual neural systems. When one individual's visual system is in a good perceptive state is arguably independent of whether the other individual's visual system is so too. This sit-

uation is modeled by our independence assumption. Thus, our results suggest that it would be optimal for the visual system to either perfectly identify a stimulus or not recognize it at all (like a specialist). This may be tied to other studies showing human participants' bimodal distribution of subjective visibility ratings (Sergent & Dehaene, 2004; Sergent, Baillet, & Dehaene, 2005). However, confidence ratings are rather continuous than bimodal (Overgaard, Rote, Mouridsen, & Ramsøy, 2006; Ramsøy & Overgaard, 2004), which may be due to a practical problem of implementing discontinuities in biological systems. There can not be a distinction between perfect identification and chance-level guessing with no intermediate representations. Instead, biological systems can only strive for somewhat in bimodal distributions. A derived hypothesis is then that the inherent advantages of independent specializations have propelled specialization of human brain areas (even before the distinctive functional specializations arose).

These considerations are heavily speculative. The theoretical results we presented must first be developed further to make more specific empirical predictions before they can be used for validations of any theory. Possible developments are to loosen the additional, technical assumptions of equal weights and binary responses. Here, we conjecture qualitatively similar patterns. These relaxations of assumptions are nevertheless worthwhile because they will allow applying our results to more applications and settings thereby increasing their empirical content and falsifiability.

5.4 Conclusion

Using the measure of mutual information is a principled way to detect aspects of human participants' responses that are overlooked by other measures such as accuracy. With it, we can diagnose methods that aim to infer how much information human participants process. This is especially important for comparative approaches where great care must be taken to not bias one side of the comparison for purely methodological reasons: When using methods that implicitly discard information, comparisons will be biased and the underlying difference occluded. We have shown that considering mutual information reveals subtleties that can substantially change interpretations about unconscious information processing and information processing in groups. Furthermore, mutual information is a determinant of group performance—beyond the accuracy of the individual group members. Thus, it is crucial to consider mutual information whenever multiple responses are combined into aggregate statistics.

Chapter 6

List of Publications and Authors' Contributions

The main chapters in this thesis are based on the following three manuscripts. Authorship contributions are declared in the corresponding tables.

1. Chapter 2: Meyen, S., Zerweck, I. A., Amado, C., von Luxburg, U., & Franz, V. H. (2021). Advancing Research on Unconscious Priming: When can Scientists Claim an Indirect Task Advantage? *Journal of Experimental Psychology: General*. Advance Online Publication. https://doi.org/10.1037/xge0001065

   Manuscript Status: Published in Journal of Experimental Psychology: General.

Author	**Scientific Ideas (in %)**	**Data Generation (in %)**	**Analysis & Interpretation (in %)**	**Paper Writing (in %)**
Sascha Meyen	30	15	80	45
Iris A. Zerweck	5	55	5	10
Catarina Amado	5	0	0	5
Ulrike von Luxburg	30	0	0	10
Volker H. Franz	30	30	15	30

2. Chapter 3: Meyen, S., Sigg, D. M. B., von Luxburg, U., & Franz, V. H. (2021). Group Decisions Based on Confidence Weighted Majority Voting. *Cognitive Research: Principles and Implications*, 6(1), 1-13.

Manuscript Status: Published in Cognitive Research: Principles and Implications.

Author	**Scientific Ideas (in %)**	**Data Generation (in %)**	**Analysis & Interpretation (in %)**	**Paper Writing (in %)**
Sascha Meyen	70	20	75	60
Dorothee M. B. Sigg	15	80	10	5
Ulrike von Luxburg	5	0	5	10
Volker H. Franz	10	0	10	25

3. Chapter 4: Meyen, S., Göppert, F., Alber, H., von Luxburg, U., & Franz, V. H. (2021) Specialists Outperform Generalists in Ensemble Classification. *arXiv preprint arXiv:2107.04381.* `https://arxiv.org/abs/2107.04381`

Manuscript Status: Published on preprint server.

Author	**Scientific Ideas (in %)**	**Data Generation (in %)**	**Analysis & Interpretation (in %)**	**Paper Writing (in %)**
Sascha Meyen	35	-	55	45
Frieder Göppert	25	-	20	5
Helen Alber	5	-	10	5
Ulrike von Luxburg	25	-	15	40
Volker H. Franz	10	-	0	5

6.1 Acknowledgements

This project is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the CRC 1233 “Robust Vision”, project number 276693517; the Institutional Strategy of the University of Tübingen (DFG, ZUK 63); and the Cluster of Excellence “Machine Learning: New Perspectives for Science”, EXC 2064/1, project number 390727645.

Chapter 7

Bibliography

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Attneave, F. (1959). *Applications of information theory to psychology: A summary of basic concepts, methods, and results.* Henry Holt.

Austen-Smith, D., & Banks, J. S. (1996). Information aggregation, rationality, and the condorcet jury theorem. *American Political Science Review*, 34–45. https://doi.org/10.2307/2082796

Ayhan, M. S., & Berens, P. (2018). Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. *Medical Imaging with Deep Learning Conference*.

Ayhan, M. S., Kuehlewein, L., Aliyeva, G., Inhoffen, W., Ziemssen, F., & Berens, P. (2020). Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis*, 101724–101724. https://doi.org/10.1016/j.media.2020.101724

Bachrach, Y., Graepel, T., Kasneci, G., Kosinski, M., & Van Gael, J. (2012). Crowd IQ: aggregating opinions to boost performance. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1* (pp. 535–542).

Bahadur, R. (1961). A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (Ed.), *Studies on Item Analysis and Prediction* (pp. 158–168). Standford University Press.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Bang, D., & Frith, C. D. (2017). Making better decisions in groups. *Royal Society Open Science*, *4*(8), 170193. https://doi.org/10.1098/rsos.170193

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y., ... Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition*, *26*, 13–23. https://doi.org/10.1016/j.concog.2014.02.002

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2004). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Bartlett, P., Freund, Y., Lee, W. S., & Schapire, R. E. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, *26*(5), 1651–1686. https://doi.org/10.1214/aos/1024691352

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2), 11–11. https://doi.org/10.1167/19.2.11

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, *5*(4), 537–550. https://doi.org/10.1109/72.298224

Birnbaum, M. H., & Diecidue, E. (2015). Testing a class of models that includes majority rule and regret theories: Transitivity, recycling, and restricted branch independence. *Decision*, *2*(3), 145–190. https://doi.org/10.1037/dec0000031

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bonnasse-Gahot, L., & Nadal, J.-P. (2008). Neural coding of categories: information efficiency and optimal population codes. *Journal of Computational Neuroscience*, *25*(1), 169–187. https://doi.org/10.1007/s10827-007-0071-5

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin*, *16*(10), 335–338. https://doi.org/10.1037/h0074554

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, *65*(3), 212 - 219. https://doi.org/10.1006/obhd.1996.0021

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology*, *40*(3), 227–259. https://doi.org/10.1006/cogp.1999.0731

Campbell, D., & Kenny, D. (1999). *Methodology in the social sciences. a primer on regression artifacts.* Guilford Press.

Chen, H., Cohn, A. G., & Yao, X. (2012). Ensemble learning by negative correlation learning. In C. Zhang & Y. Ma (Eds.), *Ensemble Machine Learning* (pp. 177–201). Springer. https://doi.org/10.1007/978-1-4419-9326-7_6

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170–178. https://doi.org/10.1016/S1364-6613(00)01476-5

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology*, *36*(1), 28–71. https://doi.org/10.1006/cogp.1998.0681

Cleeremans, A., & Tallon-Baudry, C. (2021). The function of consciousness is to generate experience. *PsyArXiv*. https://doi.org/10.31234/osf.io/jfpw2

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Colagiuri, B., & Livesey, E. (2016). Contextual cuing as a form of nonconscious learning: Theoretical and empirical analysis in large and very large samples. *Psychonomic Bulletin & Review*, *23*(6), 1996–2009. https://doi.org/10.3758/s13423-016-1063-0

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley-Interscience. https://doi.org/10.1002/047174882X

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. https://doi.org/10.1177/0956797613504966

Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science*, *329*, 47–50. https://doi.org/10.1126/science.1188595

Damian, M. F. (2001). Congruity effects evoked by subliminally presented primes: Automaticity rather than semantic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 154–165. https://doi.org/10.1037/0096-1523.27.1.154

De Condorcet, N., et al. (2014). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Cambridge University Press.

DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *32*(1-2), 12–22. https://doi.org/10.2307/2987588

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B., & Riviere, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*, 752–758. https://doi.org/10.1038/89551

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., ... Le Bihan, D. (1998, October, 8th). Imaging unconscious semantic priming. *Nature*, *395*, 597–600. https://doi.org/10.1038/26967

Dell'Acqua, R., & Grainger, J. (1999). Unconscious semantic priming from pictures. *Cognition*, *73*(1), B1–B15. https://doi.org/10.1016/S0010-0277(99)00049-9

de Zilva, D., Vu, L., Newell, B. R., & Pearson, J. (2013). Exposure is not enough: Suppressing stimuli from awareness can abolish the mere exposure effect. *PLoS ONE*, *8*(10), e77726. https://doi.org/10.1371/journal.pone.0077726

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290. https://doi.org/10.1177/1745691611406920

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*, *19*(2), 674–681. https://doi.org/10.1016/j.concog.2009.09.009

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1–15). https://doi.org/10.1007/3-540-45014-9_1

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & Van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science*, *311*(5763), 1005–1007. https://doi.org/10.1126/science.1121629

Dosher, A. B. (1998). The response–window method — some problematic assumptions: Comment on Draine and Greenwald (1998). *Journal of Experimental Psychology: General*, *127*(3), 311–317. Retrieved from https://doi.org/10.1037/0096-3445.127.3.311

Draine, S. C., & Greenwald, A. G. (1998). Replicable unconscious semantic priming. *Journal of Experimental Psychology: General*, *127*(3), 286–303. https://doi.org/10.1037/0096-3445.127.3.286

Dulaney, D. E., & Eriksen, C. W. (1959). Accuracy of brightness discrimination as measured by concurrent verbal responses and GSRs. *Journal of Abnormal and Social Psychology*, *59*, 418–423. https://doi.org/10.1037/h0040134

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological review*, *85*(5), 395–416. https://doi.org/10.1037/0033-295X.85.5.395

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158–172. https://doi.org/10.1037/0033-2909.84.1.158

Erdelyi, M. H. (1986). Experimental indeterminacies in the dissociation paradigm of subliminal perception. *Behavioral and Brain Sciences*, *9*(1), 30–31. https://doi.org/10.1017/S0140525X00021348

Eriksen, C. W. (1960). Discrimination and learning without awareness—a methodological survey and evaluation. *Psychological Review*, *67*(5), 279–300. https://doi.org/10.1037/h0041622

Eriksen, C. W., & Hake, H. W. (1955a). Absolute judgments as a function of stimulus range and number of stimulus and response categories. *Journal of Experimental Psychology*, *49*(5), 323–332. https://doi.org/10.1037/h0044211

Eriksen, C. W., & Hake, H. W. (1955b). Multidimensional stimulus differences and accuracy of discrimination. *Journal of Experimental Psychology*, *50*(3), 153–160. https://doi.org/10.1037/h0047863

Estévez, P. A., Tesmer, M., Perez, C. A., & Zurada, J. M. (2009). Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, *20*(2), 189–201. https://doi.org/10.1109/TNN.2008.2005601

Estlund, D. M. (1994). Opinion leaders, independence, and condorcet's jury theorem. *Theory and Decision*, *36*(2), 131–162. https://doi.org/10.1007/BF01079210

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Fedorov, V., Mannino, F., & Zhang, R. (2009). Consequences of dichotomization. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, *8*(1), 50–61. https://doi.org/10.1002/pst.331

Fiedler, K., & Krueger, J. I. (2012). More than an artifact: Regression as a theoretical construct. In J. I. Krueger (Ed.), *Social Judgment and Decision Making* (pp. 171–189). Psychology Press.

Finkbeiner, M. (2011). Subliminal priming with nearly perfect performance in the prime-classification task. *Attention, Perception, & Psychophysics*, *73*(4), 1255–1265. https://doi.org/10.3758/s13414-011-0088-8

Finkbeiner, M., & Palermo, R. (2009). The role of spatial attention in nonconscious processing: A comparison of face and nonface stimuli. *Psychological Science*, *20*, 42–51. https://doi.org/10.1111/j.1467-9280.2008.02256.x

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*(6), 381–391. https://doi.org/10.1037/h0055392

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280-1286. https://doi.org/10.1098/rstb.2012.0021

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends in Cognitive Sciences*, *15*(6), 272–279. https://doi.org/10.1016/j.tics.2011.04.002

Franz, V. H., & Gegenfurtner, K. R. (2008). Grasping visual illusions: Consistent data and no dissociation. *Cognitive Neuropsychology*, *25*(7), 920–950. https://doi.org/10.1080/02643290701862449

Franz, V. H., & von Luxburg, U. (2015). No evidence for unconscious lie detection: A significant difference does not imply accurate classification. *Psychological Science*, *26*(10), 1646–1648. (preprint at arXiv:1407.4240) https://doi.org/10.1177/0956797615597333

Franz, V. H., & von Luxburg, U. (2014). Unconscious lie detection as an example of a widespread fallacy in the neurosciences. *arXiv preprint arXiv:1407.4240*. https://arxiv.org/abs/1407.4240

Galesic, M., Barkoczi, D., & Katsikopoulos, K. (2018). Smaller crowds outperform larger crowds and individuals in realistic task conditions. *Decision*, *5*(1), 1–15. https://doi.org/10.1037/dec0000059

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451. https://doi.org/10.1038/075450a0

Gao, W., & Zhou, Z.-H. (2013). On the doubt about margin explanation of boosting. *Artificial Intelligence*, *203*, 1–18. https://doi.org/10.1016/j.artint.2013.07.002

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review*, *67*(6), 343–352. https://doi.org/10.1037/h0043047

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts.* Wiley.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, *5*(3-4), 303–336. http://doi.org/10.1561/105.00000092

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. https://doi.org/10.1037/0033-295X.98.4.506

Goldiamond, I. (1958). Indicators of perception: I. subliminal perception, subception, unconscious perception: An analysis in terms of psychophysical indicator methodology. *Psychological Bulletin*, *55*(6), 373–411. https://doi.org/10.1037/h0046992

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. Peninsula.

Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science*, *273*(5282), 1699–1702. https://doi.org/10.1126/science.273.5282.1699

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. *Blackwell Handbook of Judgment and Decision Making*, 177–199. https://doi.org/10.1002/9780470752937.ch9

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411–435. https://doi.org/10.1016/0010-0285(92)90013-R

Griffith, V., & Koch, C. (2014). Quantifying synergistic mutual information. In M. Prokopenko (Ed.), *Guided Self-Organization: Inception* (pp. 159–190). Springer. https://doi.org/10.1007/978-3-642-53734-9_6

Grofman, B., Owen, G., & Feld, S. L. (1983). Thirteen theorems in search of the truth. *Theory and Decision*, *15*(3), 261–278.

Hake, H. W., & Garner, W. (1951). The effect of presenting various numbers of discrete steps on scale reading accuracy. *Journal of Experimental Psychology*, *42*(5), 358–366. https://doi.org/10.1037/h0055485

Hannula, D. E., Simons, D. J., & Cohen, N. J. (2005). Imaging implicit perception: promise and pitfalls. *Nature Reviews Neuroscience*, *6*(3), 247–255. https://doi.org/10.1038/nrn1630

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. https://doi.org/10.1109/34.58871

Harder, M., Salge, C., & Polani, D. (2013). Bivariate measure of redundant information. *Physical Review E*, *87*(012130), 1–14. https://doi.org/10.1103/PhysRevE.87.012130

Hassin, R. R. (2013). Yes It Can: On the Functional Abilities of the Human Unconscious. *Perspectives on Psychological Science*, *8*, 195–207. https://doi.org/10.1177/1745691612460684

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*(2), 494–508. https://doi.org/10.1037/0033-295X.112.2.494

Hautz, W. E., Kämmer, J. E., Schauber, S. K., Spies, C. D., & Gaissmaier, W. (2015). Diagnostic performance by medical students working individually or in teams. *Journal of the American Medical Association*, *313*, 303-304. https://doi.org/10.1001/jama.2014.15770

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Herzog, M. H., Francis, G., & Clarke, A. (2019). *Understanding statistics and experimental design: How to not lie with statistics*. Springer Nature. https://doi.org/10.1007/978-3-030-03499-3

Hesselmann, G., & Moors, P. (2015). Definitely maybe: can unconscious processes perform the same functions as conscious processes? *Frontiers in Psychology*, *6*, 584. https://doi.org/10.3389/fpsyg.2015.00584

Hick, W. E. (1952). On the rate of gain of information. *Journal of Experimental Psychology*, *4*(1), 11–26. https://doi.org/10.1080/17470215208416600

Hogben, D., Pinkham, R., & Wilk, M. (1961). The moments of the non-central t-distribution. *Biometrika*, *48*(3/4), 465–468. https://doi.org/10.2307/2332772

Holender, D. (1986). Semantic activation without conscious identification in dichotic–listening, parafoveal vision, and visual masking — a survey and appraisal. *Behavioral and Brain Sciences*, *9*(1), 1-23. https://doi.org/10.1017/S0140525X00021269

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*(3), 188–196. https://doi.org/10.1037/h0056940

Irvine, E. (2013). Measures of consciousness. *Philosophy Compass*, *8*(3), 285–297. https://doi.org/10.1111/phc3.12016

Irvine, E. (2017). Explaining what? *Topoi*, *36*(1), 95–106. https://doi.org/10.1007/s11245-014-9273-4

Janssen, P. A., Thiessen, P., Klein, M. C., Whitfield, M. F., MacNab, Y. C., & Cullis-Kuhl, S. C. (2007). Standards for the measurement of birth weight, length and head circumference at term in neonates of european, chinese and south asian ancestry. *Open Medicine*, *1*(2), E74-E88.

Jensen, A. R. (1992). The importance of intraindividual variation in reaction time. *Personality and Individual Differences*, *13*(8), 869–881. https://doi.org/10.1016/0191-8869(92)90004-9

Kang, K., Shapley, R. M., & Sompolinsky, H. (2004). Information tuning of populations of neurons in primary visual cortex. *Journal of Neuroscience*, *24*(15), 3726–3735. https://doi.org/10.1523/JNEUROSCI.4272-03.2004

Kaniovski, S., & Zaigraev, A. (2011). Optimal jury design for homogeneous juries with correlated votes. *Theory and Decision*, *71*(4), 439–459. https://doi.org/10.1007/s11238-009-9170-2

Kiefer, M. (2002). The N400 is modulated by unconsciously perceived masked words: Further evidence for an automatic spreading activation account of N400 priming effects. *Cognitive Brain Research*, *13*(1), 27–39. https://doi.org/10.1016/S0926-6410(01)00085-4

Kim, C.-Y., & Blake, R. (2005). Psychophysical magic: rendering the visible 'invisible'. *Trends in Cognitive Sciences*, *9*(8), 381–388. https://doi.org/10.1016/j.tics.2005.06.012

King, J.-R., & Dehaene, S. (2014). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1641). https://doi.org/10.1098/rstb.2013.0204

Klauer, K. C., Greenwald, A. G., & Draine, S. C. (1998). Correcting for measurement error in detecting unconscious cognition: Comment on Draine and Greenwald (1998). *Journal of Experimental Psychology: General*, *127*(3), 318–319. https://psycnet.apa.org/doi/10.1037/0096-3445.127.3.318

Klein, N., & Epley, N. (2015). Group discussion improves lie detection. *Proceedings of the National Academy of Sciences*, *112*(24), 7460–7465. https://doi.org/10.1073/pnas.1504048112

Klemmer, E., & Frick, F. C. (1953). Assimilation of information from dot and matrix patterns. *Journal of Experimental Psychology*, *45*(1), 15–19. https://doi.org/10.1037/h0060868

Klotz, W., & Neumann, O. (1999). Motor activation without conscious discrimination in metacontrast masking. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 976–992. https://doi.org/10.1037/0096-1523.25.4.976

Koltchinskii, V., Panchenko, D., et al. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, *30*(1), 1–50. https://doi.org/10.1214/aos/1015362183

Koriat, A. (2012a). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. https://doi.org/10.1037/a0025648

Koriat, A. (2012b). When are two heads better than one and why? *Science*, *336*(6079), 360–362. https://doi.org/10.1126/science.1216549

Koriat, A. (2015). When two heads are better than one and when they can be worse: The amplification hypothesis. *Journal of Experimental Psychology: General*, *144*(5), 934–950. https://doi.org/10.1037/xge0000092

Koriat, A. (2017). Can people identify "deceptive" or "misleading" items that tend to produce mostly wrong answers? *Journal of Behavioral Decision Making*, *30*(5), 1066–1077. https://doi.org/10.1002/bdm.2024

Kosinski, M., Bachrach, Y., Kasneci, G., Van-Gael, J., & Graepel, T. (2012). Crowd IQ: Measuring the intelligence of crowdsourcing platforms. In *Proceedings of the 4th annual ACM web science conference* (pp. 151–160). 10.1145/2380718.2380739

Kouider, S., & Dehaene, S. (2007). Levels of processing during non–conscious perception: a critical review of visual masking. *Philosophical Transactions of*

the Royal Society B: Biological Sciences, *362*, 857–875. https://doi.org/10.1098/rstb.2007.2093

Kouider, S., & Dehaene, S. (2009). Subliminal number priming within and across the visual and auditory modalities. *Experimental Psychology*, *56*, 418–433. https://doi.org/10.1027/1618-3169.56.6.418

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, *51*(2), 181–207. https://doi.org/10.1023/A:1022859003006

Kunde, W., Kiesel, A., & Hoffmann, J. (2003). Conscious control over the content of unconscious cognition. *Cognition*, *88*(2), 223–242. https://doi.org/10.1016/S0010-0277(03)00023-4

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*(4430), 557–558. https://doi.org/10.1126/science.7352271

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., ... Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, *113*(31), 8777–8782. https://doi.org/10.1073/pnas.1601827113

Ladha, K. K. (1992). The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 617–634. https://doi.org/10.2307/2111584

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems* (pp. 6405–6416).

Lau, H. C. (2007). A higher order bayesian decision theory of consciousness. *Progress in Brain Research*, *168*, 35–48. https://doi.org/10.1016/S0079-6123(07)68004-2

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, *7*(1), 1–14. https://doi.org/10.1038/s41598-017-17876-z

Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: On the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review*, *118*(2), 379–392. https://doi.org/10.1037/a0023010

Litvinova, A., Herzog, S. M., Kall, A. A., Pleskac, T. J., & Hertwig, R. (2020). How the "wisdom of the inner crowd" can boost accuracy of confidence judgments. *Decision*(3), 183–211. https://doi.org/10.1037/dec0000119

Litvinova, A., Kurvers, R. H., Hertwig, R., & Herzog, S. M. (2019). When experts make inconsistent decisions.

https://doi.org/10.31234/osf.io/dtaz3

Liu, Y., & Yao, X. (1997). Negatively correlated neural networks can produce best ensembles. *Australian Journal of Intelligent Information Processing Systems*, *4*(3/4), 176–185.

Liu, Y., & Yao, X. (1999). Ensemble learning via negative correlation. *Neural Networks*, *12*(10), 1399–1404. https://doi.org/10.1016/S0893-6080(99)00073-8

Lu, L., Yuan, Y. C., & McLeod, P. L. (2012). Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review*, *16*(1), 54–75. https://doi.org/10.1177/1088868311417243

Luce, R. D., et al. (1986). *Response times: Their role in inferring elementary mental organization* (No. 8). Oxford University Press on Demand.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). Psychology Press.

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., ... others (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112*(12), 3835–3840. https://doi.org/10.1073/pnas.1421692112

Mamassian, P. (2016). Visual confidence. *Annual Review of Vision Science*, *2*, 459–481. https://doi.org/10.1146/annurev-vision-111815-114630

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. https://doi.org/10.1037/a0036677

Marshall, J. A., Brown, G., & Radford, A. N. (2017). Individual confidence-weighting and group decision-making. *Trends in Ecology & Evolution*, *32*(9), 636 - 645. https://doi.org/10.1016/j.tree.2017.06.004

Martins, A. C. (2006). Probability biases as bayesian inference. *Judgment and Decision Making*, *1*(2), 108–117.

Masnadi-Shirazi, H. (2013). Refinement revisited with connections to Bayes error, conditional entropy and calibrated classifiers. *arXiv preprint arXiv:1303.2517*. https://arxiv.org/abs/1303.2517

Masnadi-Shirazi, H. (2017). Combining forecasts using ensemble learning. *arXiv preprint arXiv:1707.02430*. https://arxiv.org/abs/1707.02430

Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, *5*, 1455. https://doi.org/10.3389/fpsyg.2014.01455

Mattler, U. (2003). Priming of mental operations by masked stimuli. *Perception & Psychophysics*, *65*(2), 167–187. https://doi.org/10.3758/BF03194793

Mattler, U. (2005). Inhibition and decay of motor and nonmotor priming. *Perception & Psychophysics*, *67*(2), 285–300. https://doi.org/10.3758/BF03206492

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. Psychology Press.

Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700. https://doi.org/10.1016/j.tics.2016.07.001

Merikle, P. M. (1992). Perception without awareness: Critical issues. *American Psychologist*, *47*(6), 792–795. https://doi.org/10.1037/0003-066X.47.6.792

Merikle, P. M., & Reingold, E. M. (1998). On demonstrating unconscious perception: Comment on Draine and Greenwald (1998). *Journal of Experimental Psychology: General*, *127*(3), 304–310. https://doi.org/10.1037/0096-3445.127.3.304

Meyen, S. (2016). *Relation between classification accuracy and mutual information in equally weighted classification tasks* (Unpublished master's thesis). University of Hamburg.

Meyen, S., Göppert, F., Alber, H., von Luxburg, U., & Franz, V. H. (2021). Specialists outperform generalists in ensemble classification. *arXiv preprint arXiv:2107.04381*. https://arxiv.org/abs/2107.04381

Meyen, S., Sigg, D. M., von Luxburg, U., & Franz, V. H. (2021). Group decisions based on confidence weighted majority voting. *Cognitive Research: Principles and Implications*, *6*(1), 1–13. https://doi.org/10.1186/s41235-021-00279-0

Meyen, S., Zerweck, I. A., Amado, C., von Luxburg, U., & Franz, V. H. (2020, September 17). Advancing research on unconscious priming: When can scientists claim an indirect task advantage? Retrieved from osf.io/kp59h.

Meyen, S., Zerweck, I. A., Amado, C., von Luxburg, U., & Franz, V. H. (2021). Advancing research on unconscious priming: When can scientists claim an indirect task advantage? *Journal of Experimental Psychology: General*. Advance Online Publication. https://doi.org/10.1037/xge0001065

Michel, M., Beck, D., Block, N., Blumenfeld, H., Brown, R., Carmel, D., ... Yoshida, M. (2019). Opportunities and challenges for a maturing science of consciousness. *Nature Human Behaviour*, *3*, 104–107. https://doi.org/10.1038/s41562-019-0531-8

Miller, J. (1996). The sampling distribution of d'. *Perception & Psychophysics*, *58*(1), 65–72. https://doi.org/10.3758/BF03205476

Miller, J. (2000). Measurement error in subliminal perception experiments: Simulation analyses of two regression methods. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(4), 1461–1477. https://doi.org/10.1037/0096-1523.26.4.1461

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin & Review*, *20*, 819–858. https://doi.org/10.3758/s13423-013-0404-5

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Moors, P., & Hesselmann, G. (2018). A critical reexamination of doing arithmetic nonconsciously. *Psychonomic Bulletin & Review*, *25*(1), 472–481. https://doi.org/10.3758/s13423-017-1292-x

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'bayesfactor'. *URL http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor pdf i (accessed 1006 15)*.

Morris, J. S., Öhman, A., & Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature*, *393*, 467–470. https://doi.org/10.1038/30976

Morris, J. S., Öhman, A., & Dolan, R. J. (1999). A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences of the United States of America*, *96*, 1680–1685. https://doi.org/10.1073/pnas.96.4.1680

Morrison, D., & Henkel, R. (Eds.). (1970). *The Significance Test Controversy*. Aldine Transaction.

Mudrik, L., Breska, A., Lamy, D., & Deouell, L. Y. (2011). Integration without awareness: Expanding the limits of unconscious processing. *Psychological science*, *22*(6), 764–770. https://doi.org/10.1177/0956797611408736

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. *Psychological Science*, *13*(5), 416–424. https://doi.org/10.1111/1467-9280.00474

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

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

Neth, H., Sims, C. R., & Gray, W. D. (2016). Rational task analysis: A methodology to benchmark bounded rationality. *Minds and Machines*, *26*(1-2), 125–148. https://doi.org/10.1007/s11023-015-9368-8

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*(1), 1–19. https://doi.org/10.1017/S0140525X12003214

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105–1107. https://doi.org/10.1038/nn.2886

Nitzan, S., & Paroush, J. (1982). Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, *32*(2), 289–297. https://doi.org/10.2307/2526438

Nitzan, S., & Paroush, J. (1984). The significance of independent decisions in uncertain dichotomous choice situations. *Theory and Decision*, *17*(1), 47–60. https://doi.org/10.1007/BF00140055

Nuhfer, E., Cogan, C., Fleisher, S., Gaze, E., & Wirth, K. (2016). Random number simulations reveal how random noise affects the measurements and graphical portrayals of self-assessed competency. *Numeracy: Advancing Education in Quantitative Literacy*, *9*(1). http://doi.org/10.5038/1936-4660.9.1.4

Nuhfer, E., Fleisher, S., Cogan, C., Wirth, K., & Gaze, E. (2017). How random noise and a graphical convention subverted behavioral scientists' explanations of self-assessment data: Numeracy underlies better alternatives. *Numeracy: Advancing Education in Quantitative Literacy*, *10*(1). http://doi.org/10.5038/1936-4660.10.1.4

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Computational Biology*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Okun, O., Valentini, G., & Re, M. (2011). *Ensembles in machine learning applications* (Vol. 373). Springer Science & Business Media. https://doi.org/10.1007/978-3-642-22910-7

Olbrich, E., Bertschinger, N., & Rauh, J. (2015). Information decomposition and synergy. *Entropy*, *17*(5), 3501–3517. https://doi.org/10.3390/e17053501

Olsson, H. (2014). Measuring overconfidence: Methodological problems and statistical artifacts. *Journal of Business Research*, *67*(8), 1766–1770. https://doi.org/10.1016/j.jbusres.2014.03.002

Overgaard, M., Rote, J., Mouridsen, K., & Ramsøy, T. Z. (2006). Is conscious perception gradual or dichotomous? a comparison of report methodologies during a visual task. *Consciousness and Cognition*, *15*(4), 700–708. https://doi.org/10.1016/j.concog.2006.04.002

Palfi, B., & Dienes, Z. (2020). Why Bayesian "evidence for H_1" in one condition and Bayesian "evidence for H_0" in another condition does not mean good-enough Bayesian evidence for a difference between the conditions. *Advances in Methods and Practices in Psychological Science*, *3*(3), 300–308. https://doi.org/10.1177/2515245920913019

Peirce, C. S., & Jastrow, J. (1884). On small differences in sensation. *Memoirs of the National Academy of Sciences*, *3*.

Peremen, Z., & Lamy, D. (2014). Do conscious perception and unconscious processing rely on independent mechanisms? A meta-contrast study. *Consciousness and Cognition*, *24*, 22–32. https://doi.org/10.1016/j.concog.2013.12.006

Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R. J., & Frith, C. D. (2007). How the brain translates money into force: A neuroimaging study of subliminal motivation. *Science*, *316*(5826), 904–906. https://doi.org/10.1126/science.1140459

Peters, M. A., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *Elife*, *4*, e09651. https://doi.org/10.7554/eLife.09651.001

Phillips, I. (2021a). Blindsight is qualitatively degraded conscious vision. *Psychological Review*, *128*(3), 558–584. https://doi.org/10.1037/rev0000254

Phillips, I. (2021b). Scepticism about unconscious perception is the default hypothesis. *Journal of Consciousness Studies*, *28*(3-4), 186–205.

Piasini, E., & Panzeri, S. (2019). Information theory in neuroscience. *Entropy*, *21*(1), 62. https://doi.org/10.3390/e21010062

Pishro-Nik, H. (2014). *Introduction to probability, statistics, and random processes*. Kappa Research LLC.

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America*, *24*(6), 745–749. https://doi.org/10.1121/1.1906969

Pollack, I., & Ficks, L. (1954). Information of elementary multidimensional auditory displays. *The Journal of the Acoustical Society of America*, *26*(2), 155–158. https://doi.org/10.1121/1.1907300

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Puncochar, J. M., & Fox, P. W. (2004). Confidence in individual and group decision making: When "two heads" are worse than one. *Journal of Educational Psychology*, *96*(3), 582–591. https://doi.org/10.1037/0022-0663.96.3.582

Rahnev, D., Desender, K., Lee, A. L., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., ... Zylberberg, A. (2020). The confidence database. *Nature human behaviour*, *4*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, *3*(1), 1–23. https://doi.org/10.1023/B:PHEN.0000041900.30172.e8

Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, & Psychophysics*, *80*(1), 134–154. https://doi.org/10.3758/s13414-017-1431-5

Regenwetter, M., Davis-Stober, C. P., Lim, S. H., Guo, Y., Popova, A., Zwilling, C., ... Messner, W. (2014). QTest: Quantitative testing of theories of binary choice. *Decision*, *1*(1), 2–34. https://doi.org/10.1037/dec0000007

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, *44*(6), 563–575. https://doi.org/10.3758/BF03207490

Reingold, E. M., & Merikle, P. M. (1990). On the inter-relatedness of theory and measurement in the study of unconscious processes. *Mind & Language*, *5*(1), 9–28. https://doi.org/10.1111/j.1468-0017.1990.tb00150.x

Ribeiro, M. J., Paiva, J. S., & Castelo-Branco, M. (2016). Spontaneous fluctuations in sensory processing predict within-subject reaction time variability. *Frontiers in Human Neuroscience*, *10*, 200. https://doi.org/10.3389/fnhum.2016.00200

Rothkirch, M., & Hesselmann, G. (2017). What we talk about when we talk about unconscious processing–a plea for best practices. *Frontiers in Psychology*, *8*, 835. https://doi.org/10.3389/fpsyg.2017.00835

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

Schapire, R. E., & Freund, Y. (2013). Boosting: Foundations and algorithms. *Kybernetes*. https://doi.org/10.1108/03684921311295547

Schlagbauer, B., Rausch, M., Zehetleitner, M., Müller, H. J., & Geyer, T. (2018). Contextual cueing of visual search is associated with greater subjective experience of the search display configuration. *Neuroscience of Consciousness*, *2018*(1), niy001. https://doi.org/10.1093/nc/niy001

Schmidt, T. (2002). The fingers in flight: Real–time control by visually masked color stimuli. *Psychological Science*, *13*(2), 112–118. https://doi.org/10.1111/1467-9280.00421

Schmidt, T. (2015). Invisible stimuli, implicit thresholds: Why invisibility judgments cannot be interpreted in isolation. *Advances in Cognitive Psychology*, *11*(2), 31. https://doi.org/10.5709/acp-0169-3

Schmidt, T., & Vorberg, D. (2006). Criteria for unconscious cognition: Three types of dissociation. *Perception & Psychophysics*, *68*, 489–504. https://doi.org/10.3758/BF03193692

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Seamon, J. G., Brody, N., & Kauff, D. M. (1983). Affective discrimination of stimuli that are not recognized: Effects of shadowing, masking, and cerebral laterality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(3), 544–555. https://doi.org/10.1037/0278-7393.9.3.544

Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, *8*(10), 1391–1400. https://doi.org/10.1038/nn1549

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science*, *15*(11), 720–728. https://doi.org/10.1111/j.0956-7976.2004.00748.x

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, *12*(8), 314–321. https://doi.org/10.1016/j.tics.2008.04.008

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, *24*(3), 752–775. https://doi.org/10.3758/s13423-016-1170-y

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, *37*(1), 10–21. https://doi.org/10.1109/JRPROC.1949.232969

Shapley, L., & Grofman, B. (1984). Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, *43*(3), 329–343. https://doi.org/10.1007/BF00118940

Simons, D. J., Hannula, D. E., Warren, D. E., & Day, S. W. (2007). Behavioral, neuroimaging, and neuropsychological approaches to implicit perception. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge Handbook of Consciousness* (pp. 207–250). Cambridge University Press. https://doi.org/10.1017/CBO9780511816789.010

Sklar, A. Y., Levy, N., Goldstein, A., Mandel, R., Maril, A., & Hassin, R. R. (2012). Reading and doing arithmetic nonconsciously. *Proceedings of the National Academy of Sciences*, *109*(48), 19614–19619. https://doi.org/10.1073/pnas.1211645109

Smyth, A. C., & Shanks, D. R. (2008). Awareness in contextual cuing with extended and concurrent explicit tests. *Memory & Cognition*, *36*(2), 403–415. https://doi.org/10.3758/MC.36.2.403

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational behavior and human decision processes*, *43*(1), 1–28. https://doi.org/10.1016/0749-5978(89)90055-1

Sniezek, J. A., & Henry, R. A. (1990). Revision, weighting, and commitment in consensus group judgment. *Organizational Behavior and Human Decision Processes*, *45*(1), 66–84. https://doi.org/10.1016/0749-5978(90)90005-T

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*(1), 183–203. https://doi.org/10.1037/0033-295X.108.1.183

Stahl, J., Gibbons, H., & Miller, J. (2010). Modeling single-trial LRP waveforms using gamma functions. *Psychophysiology*, *47*(1), 43–56. https://doi.org/10.1111/j.1469-8986.2009.00878.x

Stasser, G., & Abele, S. (2020). Collective choice, collaboration, and communication. *Annual Review of Psychology*, *71*(1), 589–612. 10.1146/annurev-psych-010418-103211

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*(6), 1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467

Stasser, G., & Titus, W. (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology*, *53*(1), 81–93. https://doi.org/10.1037/0022-3514.53.1.81

Stasser, G., & Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry*, *14*(3-4), 304–313. https://doi.org/10.1207/S15327965PLI1403&4_21

Stein, T., Kaiser, D., Fahrenfort, J. J., & Van Gaal, S. (2021). The human visual system differentially represents subjectively and objectively invisible stimuli. *PLoS biology*, *19*(5), e3001241. https://doi.org/10.1371/journal.pbio.3001241

Sumner, P. (2008). Mask-induced priming and the negative compatibility effect. *Experimental Psychology*, *55*(2), 133–141. https://doi.org/10.1027/1618-3169.55.2.133

ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some evidence for unconscious lie detection. *Psychological Science*, *25*(5), 1098–1105. https://doi.org/10.1177/0956797614524421

ten Brinke, L., Vohs, K. D., & Carney, D. R. (2016). Can ordinary people detect deception after all?. *Trends in Cognitive Sciences*, *20*(8), 579–588. https://doi.org/10.1016/j.tics.2016.05.012

Tononi, G. (2004). An information integration theory of consciousness. *BMC neuroscience*, *5*(1), 1–22. https://doi.org/10.1186/1471-2202-5-42

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, *23*(1), 87–102. https://doi.org/10.3758/s13423-015-0892-6

Vadillo, M. A., Linssen, D., Orgaz, C., Parsons, S., & Shanks, D. R. (2020). Unconscious or underpowered? probabilistic cuing of visual attention. *Journal of Experimental Psychology: General*, *149*(1), 160–181. https://doi.org/10.1037/xge0000632

van den Bussche, E., van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta–analysis. *Psychological Bulletin*, *135*(3), 452–477. https://doi.org/10.1037/a0015329

van Dijk, F., Sonnemans, J., & Bauw, E. (2014). Judicial error by groups and individuals. *Journal of Economic Behavior & Organization*, *108*, 224 - 235. https://doi.org/10.1016/j.jebo.2014.09.013

van Gaal, S., Ridderinkhof, K. R., Scholte, H. S., & Lamme, V. A. (2010). Unconscious activation of the prefrontal no-go network. *Journal of Neuroscience*, *30*(11), 4143–4150. https://doi.org/10.1523/JNEUROSCI.2992-09.2010

Vercammen, A., Ji, Y., & Burgman, M. (2019). The collective intelligence of random small crowds: A partial replication of Kosinski et al. (2012). *Judgment and Decision Making*, *14*(1), 91–98.

Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, *24*(1), 175–186. https://doi.org/10.1007/s00521-013-1368-0

Wallis, K. F. (2014). Revisiting francis galton's forecasting competition. *Statistical Science*, 420–424.

Wang, L., Sugiyama, M., Yang, C., Zhou, Z.-H., & Feng, J. (2008). On the margin explanation of boosting algorithms. In *COLT* (pp. 479–490).

Wang, Y., Wang, Y., Liu, P., Di, M., Gong, Y., & Tan, M. (2017). The role of representation strength of the prime in subliminal visuomotor priming. *Experimental Psychology*, *64*(6), 422–431. https://doi.org/10.1027/1618-3169/a000381

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728. https://doi.org/10.1037/0003-066X.61.7.726

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*(4), 343–364. https://doi.org/10.1037/1076-898X.2.4.343

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). McGraw–Hill Book Company. https://doi.org/10.1037/11774-000

Wójcik, M. J., Nowicka, M. M., Bola, M., & Nowicka, A. (2019). Unconscious detection of one's own image. *Psychological Science*, *30*(4), 471–480. https://doi.org/10.1177/0956797618822971

Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. (2015). Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. *PloS one*, *10*(8), e0134269. https://doi.org/10.1371/journal.pone.0134269

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting Memories* (pp. 71–94). MIT Press.

Zehetleitner, M., & Rausch, M. (2013). Being confident without seeing: what subjective measures of visual consciousness are about. *Attention, Perception, & Psychophysics*, *75*(7), 1406–1426. https://doi.org/10.3758/s13414-013-0505-2

Zerweck, I. A., Kao, C.-S., Meyen, S., Amado, C., von Eltz, M., Klimm, M., & Franz, V. H. (2021). Number processing outside awareness? systematically testing sensitivities of direct and indirect measures of consciousness. *Attention, Perception, & Psychophysics*, 1–20. https://doi.org/10.3758/s13414-021-02312-2

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1. https://doi.org/10.3389/fnins.2012.00001

Zhao, F., & Ren, Y. (2020). Revisiting contextual cueing effects: The role of perceptual processing. *Attention, Perception, & Psychophysics*, 1–15. https://doi.org/10.3758/s13414-019-01962-7

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC press.