D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.) im Fach Psychologie

Extending the Wisdom of Crowds: How to Harness the Wisdom of the Inner Crowd

---

eingereicht am Fachbereich Erziehungswissenschaft und Psychologie der Freien Universität Berlin

von **Aleksandra Litvinova, M.Sc.**

Präsident der Freien Universität Berlin:
Univ.-Prof. Dr. Günter M. Ziegler

Dekan des Fachbereichs Erziehungswissenschaft und Psychologie der Freien Universität Berlin:
Univ.-Prof. Dr. Hauke Heekeren

Promotionskommission:
Erstgutachter: Prof. Dr. Ralph Hertwig
Zweitgutachter: Prof. Dr. Rudolf Kerschreiter
Prof. Dr. Asher Koriat
Prof. Dr. Felix Blankenburg
Dr. Stefan Herzog

Tag der Einreichung: 22. November 2018

Tag der Disputation: 20. Februar 2019

Summary

Disagreement among and within individuals is not unusual. If there is, however, only one correct answer to a particular question, disagreement can lead to serious problems. Take, for example, a patient who consults with different doctors (or the same doctor on different occasions) about the malignancy of a mass in her mammogram and hears different diagnoses. Not to know which diagnosis is correct and whether the prescribed treatment is appropriate can endanger a patient's health and understandably raises concerns about the judge's competence. However, in such disagreement can lie hidden potential, a source of wisdom that becomes visible only when taking a step back to observe the disagreeing "crowd" from a distance. Such crowds often possess a fragmented and probably noisy encyclopedia of information, which is distributed over the individual members in the group. How should this fragmented knowledge be assembled into a meaningful and accurate judgment? Is there one superior strategy that consistently outperforms competing strategies? Or does the performance of each strategy depend on the statistical properties of the environment?

In the first chapter, I introduce the elementary concepts on which this dissertation draws. Crowds can be understood in many ways, of which two are relevant in this dissertation. First, a crowd can be several individuals forming a group of people, for example, a group of radiologists diagnosing the malignancy of a tumor. Second, one can think of the different opinions existing within an individual's mind as an inner crowd. For example, a single radiologist might assess the same mammogram multiple times and give different diagnoses. The main empirical body covers work on the wisdom of the inner crowd, that is, methods to harness the wisdom within an individual. In Chapter 2, I compare theoretically and empirically the performance of two strategies to improve confidence judgments of associated decisions. In the case of inconsistent judgments, should one choose the higher confidence judgment or average them? Averaging two confidence judgments consistently improves their accuracy, whereas always choosing the higher confidence judgment is risky: It can substantially harm the accuracy of confidence judgments and begins to outperform averaging only in environments where the probability of making a correct judgment is 60% or higher. Therefore, when one lacks insight into the statistical properties of an environment, the results of the presented studies suggest that averaging—due to its robustness—should be the default strategy to harness one's conflicting judgments. In the third chapter, I investigate the relationship between inconsistency in decisions, confidence judgments, and the statistical environment within expert decisions. I seek to understand when physicians change their mind, and I offer advice for people relying on expert decisions. In short, when an expert disagrees with herself, the results of this study suggest to choose the more confident decision—since the probability of making a correct decision is often 60% or higher in expert decisions. In Chapter 4, I extend the wisdom of the inner crowd to sequential diagnostic decision making and investigate cognitive dependencies between successive decisions. Existing literature suggests that gains from aggregating judgments are larger the more independent judgments are. The results so far show that the statistical properties of the environment moderate the extent of cognitive dependency processes. Future studies should investigate how such dependencies influence the accuracy of the final diagnosis. In Chapter 5, my colleagues and I review four judgment-aggregation strategies through an ecological lens. I show that there are a variety of ways to reduce uncertainty, each successful under distinct

statistical properties of the environment. If such statistical properties cannot be known, I suggest adopting two principles: (a) aggregate more judgments than fewer, and (b) use experience to adapt to the environment. Finally, in Chapter 6 I summarize the key results and point to new ideas for future research.

Taken together, the results suggest that judgment-aggregation strategies offer great potential to reduce judgment uncertainty, yet the process of doing so involves dealing with another type of uncertainty: What strategy to select in a particular environment? To investigate this and further questions I use analytical methods, computer simulations and empirical studies in different domains, ranging from mere perceptual tasks to general knowledge questions to diagnostic decisions. This work extends previous research in that it adapts and compares previous strategies and investigates them in the context of expert decisions. All in all, this sheds a different light on judgment inconsistency and shows how and when disagreement among and within individuals can be turned into a benefit.

Zusammenfassung

Meinungsdifferenzen zwischen, aber auch innerhalb, Personen findet man jeden Tag, überall und zu jeglichem Thema. Wenn es jedoch nur eine richtige Antwort auf eine Fragestellung geben kann, können Meinungsunterschiede zum Problem werden. Lässt eine Patientin zum Beispiel ihr Mammogram von verschiedenen Ärzt/innen (oder von derselben Ärztin mehrmals) auf Krebs untersuchen, kommt es durchaus vor, dass unterschiedliche Diagnosen gegeben werden. Für die Patientin kann es eine enorme Belastung sein, nicht zu wissen welche Diagnose zutrifft, und ob sie eine angemessene Behandlung bekommt. Verständlicherweise wirft dies Zweifel an der Kompetenz der Ärzt/innen auf. Schliesslich kann nur eine Diagnose stimmen. Aus einem anderen Blickpunkt betrachtet kann sich jedoch hinter diesen Unstimmigkeiten ein Potenzial—eine Art Intelligenz—verstecken. Dieses wird erst erkennbar, sobald man einen Schritt zurücktritt und den vielen unterschiedlichen Meinungen, aus der Ferne betrachtet, eine umfassende Gestalt gibt. Man könnte zum Beispiel annehmen, dass jede einzelne Meinung ein Teil eines Puzzles ist und alle, oder ein Teil der Meinungen zusammen das Bild erst vervollständigen—die "Weisheit der Vielen". Die Herausforderung liegt darin, wie man das Puzzle zusammenlegen soll. Liegen Teile dabei, die zu einem anderen Puzzle gehören? Welche Strategien stehen zur Verfügung um die Teile zusammenzufügen? Und in welchen statistischen Umgebungen führen welche Strategien zum Erfolg?

In Kapitel 1 stelle ich die grundlegenden Konzepte vor, auf die sich diese Arbeit bezieht. Die Weisheit der Vielen kann in verschiedenen Gestalten vorkommen. Zwei davon sind relevant für diese Dissertation. Erstens, kann Wissen über verschiedene Personen verteilt sein, vergleichbar zu Ärzten die unterschiedliche Diagnosen geben. Zweitens, kann Wissen auch innerhalb einer Person verteilt sein, zum Beispiel wenn ein Arzt dasselbe Mammogram zweimal evaluiert und dabei zu unterschiedlichen Diagnosen kommt. Ein Großteil dieser Arbeit untersucht, wann sich unterschiedliche Meinungen in einer Person manifestieren und vergleicht den Erfolg von verschiedenen Strategien um Meinungen zu aggregieren. Im zweiten Kapitel befasse ich mich mit subjektiven Wahrscheinlichkeitsurteilen in Entscheidungsszenarien mit zwei Alternativen. Wenn sich zwei Wahrscheinlichkeitsurteile von einer Person unterscheiden, wann sollte man Ihren Mittelwert nehmen und wann das höhere Wahrscheinlichkeitsurteil wählen? Theoretische und empirische Resultate zeigen, dass der Mittelwert zweier Wahrscheinlichkeitsurteile eine robuste Strategie ist, um ein akkurateres Wahrscheinlichkeitsurteil zu erzielen, wohingegen es eine riskante Strategie ist, sich immer auf das höhere Wahrscheinlichkeitsurteil zu verlassen. Das höhere Wahrscheinlichkeitsurteil kann Abweichung vom wahren Wert erheblich erhöhen und uebertrifft nur dann den Erfolg des Mittelwerts, wenn die objektive Wahrschenlichkeit, dass man richtig antwortet bei 60% oder höher liegt. In Kapitel 3 untersuche ich das Verhältnis zwischen inkonsistenten Entscheidungen, subjektiven Wahrscheinlichkeiten und der statistischen Umgebung in Expertenentscheidungen. Im Detail, versuche ich zu verstehen, wann Mediziner/innen ihre Meinung ändern und biete Hilfestellungen für diejenigen, die sich auf Expertenentscheidungen verlassen müssen. Kurzum, wenn ein/e Experte/in sich widerspricht, kann man sich auf die Entscheidung mit dem höheren Wahrscheinlichkeitsurteil verlassen—da in der Regel Expert/innen eine objektive Wahrscheinlichkeit richtig zu liegen von 60% erreichen oder übertreffen. In Kapitel 4 erforsche ich die Abhängigkeit zwischen aufeinanderfolgenden Teilentscheidungen in einem sequenziellen Diagnose Ver-

fahren. In der Regel lassen sich Fehler besser ausgleichen, je unabhängiger die einzelnen Entscheidungen sind. Bisher zeigen die Resultate, dass die statistische Umgebung moderiert, wie stark Teilentscheidungen in einem sequenziellen Verfahren voneinander abhängen. Künftige Studien sollten erforschen, wie sich Abhängigkeiten zwischen Teilentscheidungen auf die Richtigkeit der endgültigen Diagnose auswirken. In Kapitel 5 begutachten meine Kollegen und ich vier bekannte Strategien für das Aggregieren von Entscheidungen aus einer ökologischen Perspektive. Ich zeige, dass es eine Auswahl an Methoden gibt um Unsicherheit zu reduzieren und dass jede Methode ihre eigene Nische in einer statistischen Umgebung hat. Kennt man die Umgebung nicht, kann man zwei Prinzipien folgen: (a) Aggregieren Sie lieber mehr als weniger Urteile, und (b) verwenden Sie Feedback um sich an die Umgebung anzupassen. In Kapitel 6 fasse ich die zentralen Erkenntnisse zusammen und zeige neue Ideen für zukünftige Studien auf.

Zusammengenommen, zeigt diese Arbeit, dass man durch das Aggregieren von unterschiedlichen Urteilen Unsicherheit reduzieren kann. Jedoch impliziert die Auswahl an Strategien eine andere Unsicherheit: Welche Strategie sollte man wann anwenden? Dieser und weiteren Fragen gehe ich mit Anwendung von analytischen Methoden, Computer Simulationen und empirischen Studien auf den Grund. Ich treibe bestehende Forschung voran, indem ich bisherige Strategien erweitere, gegenüberstelle und im Kontext von Expertenentscheidungen untersuche. Im Großen und Ganzen beleuchtet diese Arbeit Meinungsunterschiede aus einem anderen Blickwinkel und weist auf, wie und wann man sie zum Guten wenden kann.

Contents

1 | General Introduction

"Organisms are algorithms, and humans are not individuals – they are 'dividuals', i.e. humans are an assemblage of many different algorithms lacking a single inner voice or a single self."

Yuval Noah Harrari, Homo Deus

Each year approximately 200,000 patients in the United States alone die from preventable medical errors (Andel, Davidow, Hollander, & Moreno, 2012); many more undergo serious harm, disability, and false treatment (Berner & Graber, 2008; Blendon et al., 2002). For patients as well as for doctors the legal and financial consequences of wrong diagnoses are vast (Andel et al., 2012) and the inconsistency in judgments between and within professionals is often identified as a principal source of the problem (Einhorn, 1974; Kahneman, Rosenfield, Gandhi, & Blaser, 2016). Several studies report inconsistency in expert judgment throughout various domains, including medicine (Kirwan, De Saintonge, Joyce, & Currey, 1983; Koran, 1975; Levi, 1989; Ullman & Doherty, 1984), clinical psychology (Little, 1961; Millimet & Greenberg, 1973), neuropsychology (Garb & Schramke, 1996), finance and management (Kahneman et al., 2016), agriculture (Trumbo, Adams, Milner, & Schipper, 1962), and weather forecasting (Lusk & Hammond, 1991; Stewart et al., 1989). Simultaneously, organizations across domains—whether in the health care sector, the judiciary, or the financial sector—expect consistency from professionals in their judgments. Disagreement among and within individuals is interpreted as a source of error, because logic requires that identical cases are evaluated identically (Kahneman et al., 2016).

Yet, in this disagreement can lie hidden potential, a source of wisdom that becomes apparent only when one steps back to observe the disagreeing crowd from a distance. Viewed from a different perspective, such crowds posses a fragmented and noisy encyclopedia of information, which is distributed over the individual members of the crowd. When aggregating this fragmented knowledge leads to the cancellation of errors and to a more accurate judgment than that of the typical—or even best—individual, we speak of the wisdom of the crowds phenomenon (Bang et al., 2014; Laan, Madirolas, & De Polavieja, 2017; Malone & Bernstein, 2015; Page, 2007; Surowiecki, 2004). One of the earliest documented aggregation strategies illustrating the wisdom of crowds effect is the majority rule, that is, choosing the option that received the most votes (Condorect, 1994). Marquis de Condorcet (1994) showed that for two-alternative choices, aggregating ever more independent judgments

can boost the accuracy of the majority vote. Two key factors influencing the success of the majority rule are the diversity of judgments (and of errors) and the average individual accuracy. If individuals make diverse judgments (i.e., diverse errors) and the average individual accuracy is higher than 0.5, that is, if the majority is correct (a "kind" environment; Hertwig, 2012), errors of the aggregated judgments will cancel each other out (Ladha, 1992, 1995). A growing body of research has investigated the potential of aggregation strategies in various domains, including medical diagnostics (Kämmer, Hautz, Herzog, Kunina-Habenicht, & Kurvers, 2017; Kurvers, De Zoete, Bachman, Algra, & Ostelo, 2018; Kurvers et al., 2016; Wolf, Krause, Carney, Bogart, & Kurvers, 2015), geopolitical and economic forecasting (Atanasov et al., 2016; Budescu & Chen, 2014; Satopää, Jensen, Mellers, Tetlock, & Ungar, 2014), and machine learning (Dutta & Bonissone, 1993). In Chapter 5 my colleagues and I review four well-known crowd rules and the ecological boundary conditions for their performance.

When thinking of the wisdom of crowds phenomenon, usually what people imagine is an obvious group of individuals solving a task. However, crowds can appear in a variety of shapes and are often invisible (Malone, 2018). Community science projects where citizens report landslides to help NASA create a global landslide catalog, online groups accumulating knowledge on Wikipedia, and the 1,000 pedestrians moving through Shibuya Crossing in Tokyo every three minutes without bumping into each other are just a few examples of intelligent and sometimes hidden crowds. This dissertation represents my work on the wisdom of two types of crowds, the known "outer" crowd and the more hidden "inner" crowd. Outer crowds are defined by groups of individuals, each expressing their opinion on a given task, for example, a group of radiologists providing individual diagnoses about the malignancy of a tumor in an x-ray. The inner crowd emerges whenever a single individual expresses diverse, maybe even conflicting opinions on a given task, for example, a radiologist making different diagnoses when assessing the same x-ray multiple times. This thesis extends the research on outer crowds to inner crowds. The main part of this work investigates how fragmented knowledge should be aggregated into a meaningful and accurate judgment. When do inner crowds arise? And how can inner crowds be used to boost accuracy? Is there a superior strategy that consistently outperforms its competitors? Or does the performance of each strategy depend on the statistical properties of the environment? In the following sections I review previous research on the wisdom of the inner crowd and highlight the open questions that I set out to answer in this dissertation.

A Brief History of the Inner Crowd

The first known demonstration of the wisdom of the inner crowd was provided in a ranking task, where individuals repeatedly arranged visually identical objects according to their weight. Averaging the rankings across individuals yielded the known wisdom of crowds effect—that is, the correlation between the judged and the actual rank of the objects increased, meaning the aggregated rankings became more accurate. Astonishingly, however, averaging ever more repeated rankings of a single individual increased the correlation to the same extent as averaging the rankings of different individuals (Stroop, 1932). Because the objects were visually

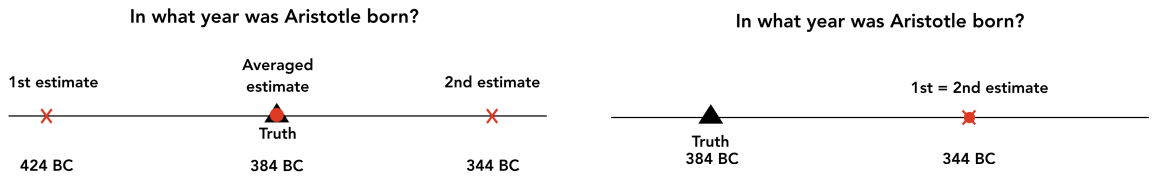
identical, individuals could not remember their previous rankings and consequently made as many diverse rankings as would different individuals.

One of the necessary conditions for the boosting effect of aggregating judgments (as well as decisions, estimates, and rankings) is a diversity of judgments, and hence a diversity of errors (Davis-Stober, Budescu, Dana, & Broomell, 2014; Herzog & Hertwig, 2009; Larrick & Soll, 2006; Page, 2007). Obviously, aggregating identical judgments cannot result in any aggregation gain. Only when judgments differ, and hence errors differ, can aggregation lead to the cancellation of errors, and therefore to a gain in accuracy. Meanwhile, a growing body of literature has shown that averaging an individual's judgments improves accuracy (for a review see Herzog & Hertwig, 2014a). But what explains the diversity in an individual's repeated judgments? One intuitive point of view is that an individual's initial judgment exhausts that person's full knowledge and any additional judgments will merely add noise (Vul & Pashler, 2008). Accordingly, initial judgments should be more accurate than consecutive judgments. Another point of view, however, is that when making judgments, individuals draw probabilistic subsamples from their knowledge base (Kersten & Yuille, 2003; Ma, Beck, Latham, & Pouget, 2006; Steyvers, Griffiths, & Dennis, 2006; Vul & Pashler, 2008). Consequently, aggregating such diverse judgments should result in the cancellation of errors. The majority of evidence for the latter proposition comes from studies averaging estimates about general knowledge quantities, such as historical dates (Herzog & Hertwig, 2009; Müller-Trede, 2011), and proportions (Fraundorf & Benjamin, 2014; Herzog & Hertwig, 2014b; Hourihan & Benjamin, 2010; Vul & Pashler, 2008), or averaging estimates about amounts in jars (Van Dolder & van den Assem, 2018).

The success of averaging can be attributed to a statistical principle: An individual's quantitative estimate can be decomposed into three parts, the true value of the quantity, *random error*, and *systematic error*. Random error is an individual's random fluctuation around the true value, while systematic error is the consistent tendency to over- or underestimate the true value. Averaging an individual's repeated estimates cancels out random error and can reduce systematic error (Herzog & Hertwig, 2009; Larrick & Soll, 2006). Take, for example, the question "In what year was Aristotle born?" Whenever two estimates bracket the true value (i.e., 384 BC), that is, when one overestimates (e.g., $estimate_1$ = 424 BC) and the other underestimates (e.g., $estimate_2$ = 344 BC) the true value, averaging them will, in this case, eliminate the absolute error ($average$ = 384 BC; $error_{average}$ = 0) and is hence more accurate than either of the two estimates (Figure 1a). When two estimates fall on the same side of the truth, that is, when both estimates over- or underestimate the true value, averaging will be at least as accurate as randomly choosing one of the two (Figure 1a; for a comparison of averaging vs. choosing, see Soll & Larrick, 2009).

Dialectical Bootstrapping: Facilitating Diversity Within an Individual's Judgments

What influences the amount of diversity within an individual's judgments? Increasing the time delay by three weeks between an individual's repeated judgments has been shown to reduce dependency between errors and hence to increase accuracy more than asking for an immediate repeated judgment (Vul & Pashler, 2008).



(a) Bracketing. First and second estimates bracket the true value. Averaging both estimates eliminates the error in this case and is more accurate than randomly choosing between one of the estimates.

(b) No bracketing. First and second estimates underestimate the true value by the same amount, hence averaging does not cancel the error.

Figure 1. Power of averaging.

Furthermore, individuals with lower working memory span have been shown to produce more independent judgments, hence yielding higher averaging gains than individuals with higher working memory span. The advantage of a lower memory span results from smaller samples drawn from the same knowledge base being necessarily less dependent than bigger samples (Hourihan & Benjamin, 2010). But what if there is not enough time to wait for a second judgment, or if an individual simply has a high memory span? How else could the dependency within an individual's repeated judgments be reduced?

Herzog and Hertwig (2009, 2014b) proposed *dialectical bootstrapping* as a general mental tool to harness the wisdom of the inner crowd, where dialectical refers to the Hegelian dialectical method of creating a thesis (first estimate), an antithesis (dialectical estimate), and a synthesis (aggregation) and bootstrapping refers to the saga of Baron Münchhausen, who escaped from a swamp by pulling himself up by his own bootstraps. Accordingly, any technique that promotes the use of different, nonredundant knowledge and reduces the influence of the first estimate falls within this framework. One implementation of dialectical bootstrapping is the "consider-the-opposite" technique, whereby individuals are encouraged to generate a dialectical estimate by thinking about reasons why their first estimate was wrong. Herzog and Hertwig (2009, 2014b) demonstrated that dialectical estimates had higher bracketing rates than simply repeated estimates (i.e., without dialectical instructions), and therefore averaging dialectical estimates resulted in higher accuracy than averaging simply repeated estimates. In other words, instead of waiting for three weeks, an individual can instantaneously benefit from her inner crowd by actively generating a plausible dialectical second estimate. Do individuals aggregate and combine their judgments without being explicitly instructed to do so? Individuals are more likely to combine their judgments when instructed to contradict themselves and with growing disagreement between their judgments (Herzog & Hertwig, 2014b).

So far research has shown that individuals do not always agree with themselves, and that they can use this disagreement to their advantage by aggregating their repeated or dialectical estimates. However, evidence so far is constrained to quantitative estimates. Could the wisdom of the inner crowd also be harnessed in categorical decision making tasks, such as in the medical diagnostic task of detecting breast cancer? How often and under what conditions do individuals disagree with themselves? And what strategies can be used to aggregate categorical decisions?

Outline of the Dissertation

The main empirical part of this dissertation extends the research on the wisdom of the inner crowd to categorical decisions and their associated confidence judgments. Each chapter is being or has been prepared for publication and can thus stand alone.[1] In Chapter 2, I investigate how repeated confidence judgments in two-alternative choices can be aggregated. Two strategies, *averaging* two confidence judgments (Ariely et al., 2000) and *maximizing*, that is, choosing the judgment with the higher confidence (Koriat, 2012), have previously been proposed but never compared against each other. In Chapter 2, I investigate theoretically and empirically which strategy performs well in particular environments. Is there a superior strategy that consistently outperforms competing strategies, or does the success of each strategy depend on the statistical environment? If higher confidence in one's decision is associated with higher accuracy, maximizing should improve accuracy. However, if the relationship between confidence and accuracy is not described by a monotonically increasing function, then maximizing can potentially harm accuracy.

Chapter 3 looks at what causes experts to change their mind and provides advice for those who have to rely on expert decisions. More precisely, I investigate the relationship between inconsistency in expert decisions, confidence judgments, and the statistical environment. Do experts change their mind more frequently when they are more likely to be wrong than correct? Or is it something about the statistical environment that drives inconsistency? The applied paradigm is similar to that outlined in Chapter 2. I stay within the realm of two-alternative choices and their associated confidence judgments.

Chapter 4 broadens the wisdom of the inner crowd research to sequential diagnostic decision making. In sequential diagnostic procedures, individuals make consecutive subdecisions before arriving at a final decision. The three-point checklist of dermoscopy is one example of a sequential procedure applied in the field of dermatology (Zalaudek et al., 2006). Radiologists first assess the presence of three cues, one at a time, and then make a final judgment based on the number of present cues. However, studies have shown that the order of evidence can bias one's final diagnosis (Rebitschek, Bocklisch, Scholz, Krems, & Jahn, 2015). Since independent and diverse judgments are one of the key factors driving aggregation gains, Chapter 4 addresses the questions of (i) whether sequential diagnostic procedures induce dependencies between subdecisions in a diagnostic sequence (taking into account the statistical properties of the environment), and (ii) whether a different procedure can reduce such dependencies.

In Chapter 5, takes a broader perspective and explores four well-known aggregation strategies for outer crowds (i.e., the opinions of several individuals) through the lens of ecological rationality. In what statistical environments do those strategies perform well? And how should one proceed if there is little information about the statistical environment? Finally, in Chapter 6, I conclude with a summary of the findings from Chapters 2–5 and provide ideas for future research.

[1] This is not a cumulative, publication–based dissertation but follows that form.

References

Andel, C., Davidow, S. L., Hollander, M., & Moreno, D. A. (2012). The economics of health care quality and medical errors. *Journal of Health Care Finance*, *39*(1), 39–50.

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., ... Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*, 130–147. doi: 10.1037/1076-898X.6.2.130

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., ... Mellers, B. (2016). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, *63*(3), 691–706. doi: 10.1287/mnsc.2015.2374

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

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5 Suppl), S2–S23. doi: 10.1016/j.amjmed.2008.01.001

Blendon, R. J., DesRoches, C. M., Brodie, M., Benson, J. M., Rosen, A. B., Schneider, E., ... Steffenson, A. E. (2002). Views of practicing physicians and the public on medical errors. *New England Journal of Medicine*, *347*(24), 1933–1940. doi: 10.1056/NEJMsa022151

Budescu, D. V., & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280. doi: 10.1287/mnsc.2014.1909

Condorect, N. C. (1994). Essay on the application of probability analyses to decisions returned by a plurality of people. In *Condorcet: Foundations of social choice and political theory* (pp. 11–36). Brookfield,VT: Edward Elgar. (Original work published 1785)

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79–101. doi: 10.1037/dec0000004

Dutta, S., & Bonissone, P. P. (1993). Integrating case-and rule-based reasoning. *International Journal of Approximate Reasoning*, *8*(3), 163–203.

Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology*, *59*(5), 562–571.

Fraundorf, S. H., & Benjamin, A. S. (2014). Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language*, *71*(1), 17–38. doi: 10.1016/j.jml.2013.10.002

Garb, H. N., & Schramke, C. J. (1996). Judgment research and neuropsychological assessment: A narrative review and meta-analyses. *Psychological Bulletin*, *120*(1), 140–153. doi: 10.1037/0033-2909.120.1.140

Hertwig, R. (2012). Tapping into the wisdom of the crowd—with confidence. *Science*, *336*(6079), 303–304.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237. doi: 10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2014a). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506. doi: 10.1016/j.tics.2014.06.009

Herzog, S. M., & Hertwig, R. (2014b). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 218–232. doi: 10.1037/a0034054

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 1068–1074. doi: 10.1037/a0019694

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*, *94*(10), 38–46.

Kämmer, J. E., Hautz, W. E., Herzog, S. M., Kunina-Habenicht, O., & Kurvers, R. H. J. M. (2017). The potential of collective intelligence in emergency medicine: Pooling medical students' independent decisions improves diagnostic performance. *Medical Decision Making*, *37*(6), 715–724. doi: 10.1177/0272989X17696998

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*(2), 150–158. doi: 10.1016/S0959-4388(03)00042-4

Kirwan, J., De Saintonge, D. C., Joyce, C., & Currey, H. (1983). Clinical judgment in rheumatoid arthritis. i. rheumatologists' opinions and the development of 'paper patients'. *Annals of the Rheumatic Diseases*, *42*(6), 644–647.

Koran, L. M. (1975). The reliability of clinical methods, data and judgments. *New England Journal of Medicine*, *293*(14), 695–701. doi: 10.1056/NEJM197510022931405

Koriat, A. (2012). When are two heads better than one and why? *Science*, *336*(6079), 360–362. doi: 10.1126/science.1216549

Kurvers, R. H. J. M., De Zoete, A., Bachman, S. L., Algra, P. R., & Ostelo, R. (2018). Combining independent

decisions increases diagnostic accuracy of reading lumbosacral radiographs and magnetic resonance imaging. *PloS ONE*, *13*(4), e0194128. doi: 10.1371/journal.pone.0194128

Kurvers, R. H. J. M., 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 of the United States of America*, *113*(31), 8777–8782. doi: 0.1073/pnas.1601827113

Laan, A., Madirolas, G., & De Polavieja, G. G. (2017). Rescuing collective wisdom when the average group opinion is wrong. *Frontiers in Robotics and AI*, *4*(56), 358–366. doi: 10.3389/frobt.2017.00056

Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 617–634.

Ladha, K. K. (1995). Information pooling through majority-rule voting: Condorcet's jury theorem with correlated votes. *Journal of Economic Behavior & Organization*, *26*, 353–372.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. doi: 10.1287/mnsc.1050.0459

Levi, K. (1989). Expert systems should be more accurate than human experts: evaluation procedures from human judgement and decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*(3), 647–657. doi: 10.1109/21.31070

Little, K. B. (1961). Confidence and reliability. *Educational and Psychological Measurement*, *21*(1), 95–101.

Lusk, C. M., & Hammond, K. R. (1991). Judgment in a dynamic task: Microburst forecasting. *Journal of Behavioral Decision Making*, *4*(1), 55–73. doi: 10.1002/bdm.3960040105

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. doi: 10.1038/nn1790

Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together*. New York, NY: Little, Brown.

Malone, T. W., & Bernstein, M. S. (2015). *Handbook of collective intelligence*. Cambridge, MA: MIT Press.

Millimet, C. R., & Greenberg, R. P. (1973). Use of an analysis of variance technique for investigating the differential diagnosis of organic versus functional involvement of symptoms. *Journal of Consulting and Clinical Psychology*, *40*(2), 188–195. doi: 10.1037/h0034568

Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. *Judgment and Decision Making*, *6*(4), 283–294.

Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Rebitschek, F. G., Bocklisch, F., Scholz, A., Krems, J. F., & Jahn, G. (2015). Biased processing of ambiguous symptoms favors the initially leading hypothesis in sequential diagnostic reasoning. *Experimental Psychology*, *62*(5), 287–305. doi: 10.1027/1618-3169/a000298

Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *The Annals of Applied Statistics*, *8*(2), 1256–1280. doi: 10.1214/14-AOAS739

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805. doi: 10.1037/a0015145

Stewart, T. R., Moninger, W. R., Brady, R. H., Merrem, F. H., Stewart, T. R., & Grassia, J. (1989). Analysis of expert judgment in a hail forecasting experiment. *Weather and Forecasting*, *4*(1), 24–34. doi: 10.1175/1520-0434(1989)004<0024:AOEJIA>2.0.CO;2

Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, *10*(7), 327–334. doi: 10.1016/j.tics.2006.05.005

Stroop, J. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology*, *15*(5), 550–562. doi: 10.1037/h0070482

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Garden City, NY: Doubleday.

Trumbo, D., Adams, C., Milner, M., & Schipper, L. (1962). Reliability and accuracy in the inspection of hard red winter wheat. *Cereal Science Today*, *7*, 62–71.

Ullman, D. G., & Doherty, M. E. (1984). Two determinants of the diagnosis of hyperactivity: The child and the clinician. *Advances in Developmental & Behavioral Pediatrics*, *5*, 167–219.

Van Dolder, D., & van den Assem, M. J. (2018). The wisdom of the inner crowd in three large natural experiments. *Nature Human Behaviour*, *2*(1), 21–26. doi: 10.1038/s41562-017-0247-6

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. doi: 10.1111/j.1467-9280.2008.02136.x

Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. J. M. (2015). Collective intelligence meets medical decision-making: The collective outperforms the best radiologist. *PloS ONE*, *10*(8), e0134269. doi: 10.1371/journal.pone.0134269

Zalaudek, I., Argenziano, G., Soyer, H. P., Corona, R., Sera, F., Blum, A., ... The Dermoscopy Working Group (2006). Three-point checklist of dermoscopy: An open internet study. *British Journal of Dermatology*, *154*(3), 431–437. doi: 10.1111/j.1365-2133.2005.06983.x

2 | How the "Wisdom of the Inner Crowd" Can Boost Accuracy of Confidence Judgments

Litvinova, A., Herzog, S.M., Kall, A.A., Pleskac, T.J. & Hertwig, R.

Abstract

The *wisdom-of-crowds* effect describes how aggregating judgments of multiple individuals can lead to a more accurate judgment than that of the typical—or even best—individual. We investigated when individuals can avail themselves of the wisdom of their "inner crowd" to improve the quality of their confidence judgments by either (a) *averaging* their two confidence judgments or (b) selecting the higher of the two (i.e., *maximizing*). In a simulation analysis based on a signal detection model of confidence, we investigated how the "kindness" versus "wickedness" of items (i.e., the degree to which the majority of people chooses the correct or wrong answer) and the redundancy of the two confidence judgments (made by the same person) affect the performance of averaging and maximizing. Simulation and analytical results show that irrespective of the type of item, averaging consistently improves confidence judgments, but maximizing is risky: It outperformed averaging only once items were answered correctly 60% of the time or more. All effects were smaller the higher the redundancy between confidence judgments. We investigated the relevance of these effects in three empirical datasets since a person's actual confidence judgments are redundant (median correlations ranged between .5 and .85). Averaging two confidence judgments from the same person was superior to maximizing, with Cohen's *d*'s effect sizes ranging from 0.67–1.44. As people typically have no insight about the wickedness of the individual item, our results suggest that averaging—due to its robustness—should be the default strategy to harness one's conflicting confidence judgments.

Keywords: judgments under uncertainty; judgment aggregation; dialectical bootstrapping; wisdom of the inner crowd; confidence judgments

---

Data and scripts for reproduction can be found at Open Science Framework. | Acknowledgments: We thank Anita Todd for editing the manuscript, and the Swiss National Science Foundation for a grant to the second and fifth author (100014_129572/1).

Introduction

Among many psychologists and economists, confidence judgments have a bit of a "bad boy" persona (Griffin & Brenner, 2004). Extant research has claimed that subjective confidence judgments violate coherence norms of rationality (Kahneman & Tversky, 1982) and do not reliably reflect people's actual decision accuracy (D. D. P. Johnson & Fowler, 2011; Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982; Sniezek, Paese, & Switzer III, 1990). Notwithstanding this notorious reputation (but see Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, Winman, & Olsson, 2000; Pleskac & Busemeyer, 2010), confidence is one of the most important correlates of acts of judgment and decision. In numerous areas of real-world decision making, such as intelligence service (Betts, 1978; Mandel & Barnes, 2014; Mellers et al., 2014), eyewitness reports (Wixted, Mickes, Dunn, Clark, & Wells, 2016), the stock market, and medical diagnostics (Berner & Graber, 2008), people cannot help but rely on confidence judgments to assess the accuracy of decisions or the likelihood of an event to happen. That is, people often treat confidence as a cue whether to act on a decision or whether they should consult additional information. The accuracy of confidence judgments is thus key.

The accuracy of confidence judgments has been well studied often showing people are overconfident and unreliable (for a review, see Arkes, 2001; McClelland & Bolger, 1994; Moore, Tenney, & Haran, 2015). Some have argued that this miscalibration does not reside in the decision maker's cognition but in the item-sampling process: Representative samples of general knowledge items do not lead to miscalibrated confidence but selectively sampled items do (Dhami, Hertwig, & Hoffrage, 2004; Gigerenzer et al., 1991; Juslin et al., 2000). Other researchers attempted to improve the quality of confidence judgments using various techniques, mostly focusing on how to elicit and improve the very first judgment a person makes (Arkes, 2001). For example, having people consider evidence inconsistent with their current belief can reduce overconfidence (Koriat, Lichtenstein, & Fischhoff, 1980). Relatedly, considering alternative outcomes and explanations can reduce bias in confidence judgments (Hirt & Markman, 1995). Other researchers attempted to improve the quality of confidence judgments by post-processing them statistically (Baron, Mellers, Tetlock, Stone, & Ungar, 2014; Satopää et al., 2014).

We took an entirely different approach to improving confidence judgments, capitalizing on the fact that sometimes people sit between a rock and a hard place, and struggle with conflicting opinions they simultaneously contemplate. As a result, people can experience an inner crowd made up of multiple, perhaps sometimes conflicting judgments about the same problem. Previous work has shown that there may be a wisdom to this inner crowd in that people can use it to inform and improve their judgments (Herzog & Hertwig, 2009, 2014b; Vul & Pashler, 2008; for a review see Herzog & Hertwig, 2014a). In this paper, we sought to understand how this *wisdom of the inner crowd* might extend to confidence judgments. We considered two strategies for harnessing the wisdom of the inner crowd: (a) Follow the highest confidence judgment (adapted from the maximum-confidence-slating technique; Koriat, 2012b), which we call *maximizing*; and (b) average one's repeated confidence judgments (Ariely et al., 2000), which we call *averaging*.

In the following, we introduce the notion of the wisdom of the crowd and how it can be applied within one's

own mind. We then discuss maximizing and averaging—both strategies representing two hitherto unconnected lines of research (Ariely et al., 2000; Koriat, 2012a)—and evaluate their potential strengths and weaknesses using a simulation and an analytical approach. We then report analyses of these strategies and their potential to boost the accuracy of confidence judgments across three empirical datasets, with two stemming from published studies and one from a new study.

The Wisdom of the (Inner) Crowd

The *wisdom-of-crowds* effect (Larrick, Mannes, & Soll, 2012; Surowiecki, 2004) describes the phenomenon that aggregating independent judgments of multiple individuals with diverse knowledge sources can lead to a more accurate judgment than that of the typical—or even best—individual by canceling out opposing errors (Larrick & Soll, 2006). Similarly, people can store diverse, perhaps even conflicting pieces of information regarding the same problem but may often rely only on a subsample of that information to arrive at a judgment at any point in time. Therefore, if they probe their knowledge again, sampling anew, they can arrive at a slightly or sometimes even drastically different judgment (Hourihan & Benjamin, 2010; Koriat, 2012a; Lewandowsky, Griffiths, & Kalish, 2009; Steyvers, Griffiths, & Dennis, 2006; Vul & Pashler, 2008). This suggests that averaging an individual's repeated quantitative estimates may result in the cancellation of both systematic biases in the sampled knowledge and unsystematic error, leading to improved estimates. Indeed, averaging an individual's repeated quantitative estimates improves accuracy (for a review see Herzog & Hertwig, 2014a), but the size of this accuracy gain depends on how correlated an individual's repeated judgments are. The accuracy can be further enhanced by increasing the time between two repeated estimates (Van Dolder & van den Assem, 2018; Vul & Pashler, 2008; but see Steegen, Dewitte, Tuerlinckx, & Vanpaemel, 2014), as well as actively encouraging an individual to approach the same question from a different angle to reduce error redundancy (Herzog & Hertwig, 2009, 2014b).

So far research on the wisdom of this *inner crowd* phenomenon—judgment aggregation within one person relative to aggregation across people—has primarily focused on improving the estimates pertaining to objective quantities, but not on how aggregation changes a person's uncertainty or confidence. Going beyond this past focus, we here present a comprehensive analysis of when and how two different ways of harnessing the potential wisdom of the inner crowd (Herzog & Hertwig, 2014a)—maximizing or averaging individual's multiple and possibly conflicting confidence judgments—improve a person's final confidence in her decision.

Maximizing builds on the result that typically the higher a person's confidence in a decision, the more likely that decision is accurate (see, e.g., Baranski & Petrusic, 1998; Dougherty, 2001; Garrett, 1922; D. M. Johnson, 1939; Kurvers et al., 2016; Nelson & Narens, 1990; Pleskac & Busemeyer, 2010; Vickers, 1979; Yaniv, Yates, & Smith, 1991; Yu, Pleskac, & Zeigenfuse, 2015). As a result, confidence can serve as a cue to the accuracy of a decision or forecast. From this perspective, when faced with the choice between two self-generated confidence judgments one could maximize and select the higher confidence judgment and its decision. Alternatively, however, one could start with the argument that two confidence judgments reflect different, possibly nonredundant

pieces of information and therefore averaging an individual's two confidence judgments is likely to result in the most accurate confidence judgment (Wallsten, Budescu, Erev, & Diederich, 1997; Wallsten & Diederich, 2001). Still another rationale is that the first judgment represents a person's best effort and additional judgments at best represent noisy, degraded versions of it (Vul & Pashler, 2008) and at worst add systematic error. In our analyses, we used a person's first confidence judgment as a benchmark and compared the performance of averaging and maximizing to a "one-and-done" policy. In the following, we review how maximizing and averaging have been investigated in previous research and introduce two crucial factors that moderate the success of both strategies.

Past research has considered a strategy similar to maximizing. Koriat (2012b) and Bang et al. (2014) investigated the effect of choosing the decision with the highest confidence (i.e., maximum confidence slating; MCS)—across and within individuals—on the accuracy of decisions, but not on the accuracy of confidence judgments. MCS did improve decision accuracy, however, only for what might be called "kind" items (Hertwig, 2012; Koriat, 2012b), that is, items for which the majority agreed on the correct answer. In contrast, for "wicked" items where the majority agreed on the *wrong* answer, the use of MCS impaired decision accuracy because the most confident decision was more likely to be wrong than the less confident decision. To illustrate, a wicked item could be "Which city is the capital of Australia: (a) Canberra or (b) Sydney?", where the majority of, for example, European citizens would answer "Sydney" because it is the more popular city. Koriat (1976, 2008, 2012a) explained this finding with the conjecture that an individual's confidence is based on an assessment of how clearly a set of sampled cues agrees with the selected response. Assuming some convergence among the population of respondents in terms of the cues in their knowledge base, this implies that there will be a relationship between an individual's confidence in her or his decision and how large the majority of people is who select that particular answer, a relationship that Koriat (2008) referred to as the *consensuality principle*.

Yet if not only the decision but also confidence is evaluated, MCS specifies which decision but not which of two possible states of confidences is more appropriate. One natural extension of the MCS strategy to confidence judgments is to assume that in light of multiple confidence judgments a person generated, the highest confidence judgment is the most accurate presumably because it based on the most coherent evidence. This is what we here refer to as maximizing. However, if confidence tracks consensuality and not accuracy, as suggested by Koriat (2012a), the effects of maximizing on the quality of confidence will be similar to the effects MCS on the accuracy of decisions. That is, it will improve the quality for kind items but impair the quality for wicked items. If this is the case, then maximizing will yield progressively worse results as the wickedness of the items increases.

Past research has investigated the effect of *averaging* confidence judgments across and *within* individuals (Ariely et al., 2000). Specifically, Ariely et al. (2000) investigated the effects of averaging on different aspects of accuracy, such as how well confidence judgments discriminate between correct and wrong decisions (i.e., *resolution*) and how well subjective confidence judgments correspond to objective probabilities (i.e., *calibration*). In general, averaging confidence judgments across or within individuals improves the overall quality

of confidence judgments. However, the benefits of averaging and its effects on different aspects of accuracy depend on the redundancy in the knowledge sources underlying confidence judgments (Erev, Wallsten, & Budescu, 1994; Wallsten et al., 1997). When the knowledge sources underlying the aggregated judgments are distinct, averaging improves the ability of confidence judgments to discriminate between correct and wrong decisions (i.e., *resolution*) but compromises the correspondence between subjective and objective probabilities (i.e., *calibration*), whereas under shared knowledge sources, averaging solely improves calibration by canceling out random error (Ariely et al., 2000; Wallsten & Diederich, 2001).

How do averaging and maximizing confidence judgments perform in a competition against each other? Relatedly, which strategy promises better results assuming that individuals lack insight into whether they face a kind or a wicked item? We investigated these questions primarily in the context of judgmental tasks (Laughlin, 1980; Laughlin & Ellis, 1986) where (simulated or actual) participants were asked to rate their confidence either in their choice or in a given event (e.g., "Sofia is the capital of: (a) Romania or (b) Bulgaria?"). Regardless of which confidence rating they gave, in all tasks our participants responded to each question twice and thus provided confidence judgments twice. Judgmental tasks differ from intellective tasks in that the latter are tasks in which the correctness of the solution can be demonstrated at the time of deliberation (e.g., mathematical tasks), whereas in judgmental tasks this correctness cannot be demonstrated online (Laughlin, 1980; Laughlin & Ellis, 1986). Forecasting a future event is the quintessential judgmental task because the outcome is not known at the time of judgment.

To understand the important influence of both the kindness of the environment and the redundancy in knowledge sources, we began our investigation by conducting a simulation study based on a signal detection model of confidence (Ferrell & McGoey, 1980; Gu & Wallsten, 2001) and an analytical model. To the best of our knowledge, in the context of the wisdom of the inner crowd, we here present the first systematic study of the boundary conditions for the success of averaging and maximizing and delineate under which conditions one strategy would have an edge over the other. Subsequently, we examine whether the analytical and simulation insights hold up in actual, empirical confidence judgments. To this end, we analyzed data from three empirical studies (two reanalyses of previously published studies and one new study), taking into account the environmental structure and correlation of confidence judgments as a proxy for the redundancy of knowledge sources underlying both judgments.

A Simulation Study of Averaging and Maximizing Confidence Judgments

We conducted a simulation study to gain insights into how the statistical structure of the knowledge environment affects the accuracy of individual confidence judgments and that of averaging and maximizing two confidence judgments. To this end, we manipulated the probability $p(C)$ [.1, .2, . . . , .9] of correctly choosing between two options and created for each value of $p(C)$ a corresponding environment consisting of many decisions based on that value of $p(C)$. Using these environments, we generated two confidence judgments per item, while systematically varying the redundancy between the knowledge sources underlying the repeated

confidence judgments from the same individual (expressed as a correlation r [0, .25, .5, .75]). By orthogonally varying the values of $p(C)$ and r, we thus created 36 different environments in total. As a result, the simulation analysis illustrates the joint effects of the kindness of the environment and the dependency in knowledge sources on the accuracy of averaging and maximizing confidence judgments. All scripts to reproduce the simulation can be found at: https://osf.io/b3f6d/?view_only=22b543c3ab3f4943af67b5c4842127d5

Methods

To systematically manipulate the kindness across environments, we constructed different environments, where within each of them all items had an identical probability $p(C)$ of being answered correctly: .1, .2, ..., or .9. [1] We adopted the framework of signal detection theory introduced by Ferrell and McGoey (1980, their 2AFC(HR) model) and further developed by Gu and Wallsten (2001) to simulate confidence judgments based on an item's value of $p(C)$. This signal detection theory model quantifies the ability of confidence judgments to discriminate between correct (signal plus noise) and incorrect decisions (noise), where the mean of the signal distribution is typically higher than that of the noise distribution. The sensitivity index, or d', is a measure of the separation of those means, where a higher d' indicates better discrimination ability.

For each item in each environment, we generated *two* confidence judgments, corresponding to the first and second confidence judgment of a simulated individual. To this end, we extended the signal detection theory framework of confidence (Ferrell & McGoey, 1980; Gu & Wallsten, 2001) by replacing the two respective *univariate* normal distributions for signal and noise trials with two *bivariate* normal distributions. This allowed us to model the redundancy of two confidence judgments. To create subjective intensities for first and second confidence judgments, we sampled *one* observation from either the signal or the noise distribution. Whether the observation was drawn from the signal or the noise distribution was determined by drawing either 1 or 0 from a Bernoulli distribution where the probability of success equaled the $p(C)$ value of the current item. The observation's value along the first dimension (x_1) corresponded to the subjective intensity of the first judgment and its value along the second dimension (x_2) corresponded to the subjective intensity of the second judgment. The signal distribution was set to have a bivariate mean of $\mu_{1,2}^{signal} = \frac{d'}{2}$ and the noise distribution of $\mu_{1,2}^{noise} = -\frac{d'}{2}$; the standard deviations of both distributions along both dimensions (x_1 and x_2) were all set to $\sigma = 1$. To determine d' for an item, we transformed the $p(C)$ value into a d' value using: $d' = \sqrt{2}\Phi(p(C))$, where Φ is the inverse of the standard normal cumulative distribution function.

To simulate different levels of dependency between the knowledge sources used for first and second judgments, we varied the correlation r in the covariance matrix underlying both bivariate distributions using the values $0, .25, .5,$ and $.75$ (i.e., we assumed that the dependency within the signal and the noise distribution is the same).

[1] We also created heterogeneous environments, where the probability $p(C)$ of being answered correctly differed across items (modeled as beta distributions). The qualitative conclusions from these additional simulations were fully in line with those of the simulations using homogeneous environments (see the Appendix A3, subsection "Heterogeneous Environments").

Finally, to translate the subjective, latent intensities into overt confidence judgments, we followed Ferrell and McGoey (1980) and Gu and Wallsten (2001) and chose a vector of 11 response categories of subjective probability judgments [0, .1, .2, . . . , 1.0] and mapped the subjective intensities onto those discrete response categories. An optimization algorithm determined the location of the category boundaries, ensuring that the confidence judgments were roughly calibrated for medium difficulty items (i.e., $d' = 1.4$). [2] The resulting confidence judgments represented the belief in being correct on a full-range probability scale. Confidence judgments that fell below 50% thus imply that the belief in being correct was higher for the opposite decision.

Results

Overall accuracy. To assess the overall accuracy of confidence judgments, we calculated the mean probability, or Brier, score (Brier, 1950):

$$PS = \frac{1}{N}\sum_{i=1}^{N}(o_i - f_i)^2\,,$$

which measures the mean squared deviation between the confidence judgments (f_i) that event o_i will happen and the actual event o_i (i.e., $o_i = 1$ if o_i happened vs. $o_i = 0$ if o_i did not happen) for N items. Zero is the best possible score and 1 the worst possible. Randomly choosing between two options and then assigning .5 confidence to each decision would yield a score of .25.

Because the first and second confidence judgments perform equally well by construction, we compared the performance of averaged and maximized confidence judgments only against that of first confidence judgments. Figure 1 shows the Brier score as a function of the probability of being correct ($p(C)$) and the redundancy in the knowledge sources (r). As expected by the design of the simulation, as $p(C)$ increased, Brier scores decreased for first, averaged, and maximized confidence judgments, reflecting the fact that as items became more kind, confidence judgments became more accurate.

Comparing averaged to first judgments, averaging improved the Brier score in all environments—even in wicked environments (i.e., $p(C) < .5$). For example, in $r = 0$ and $p(C) = 0.2$ (Figure 1, left-most panel), averaging improved the Brier score by .028 points. The greatest gains from averaging were concentrated in the central range of $p(C)$ [.4, .7], an improvement of .03 points (for $r = 0$). When first and second confidence judgments became more similar (i.e., as redundancy, r, increased), these differences decreased and the Brier score of averaged judgments converged to that of first judgments—illustrating that diversity in judgments is a key requisite for the wisdom-of-crowds effect. In stark contrast to averaging, the effects of maximizing

[2] Somewhat counterintuitively, perfect calibration is only possible for medium difficulty levels (i.e., $d' \approx 1.4$), but is not even possible in principle for difficult and very easy decisions (Ferrell & McGoey, 1980; Gu & Wallsten, 2001). We therefore optimized the category boundaries for $d' = +1.4$ once and then used this one fixed set of boundaries throughout the simulation. This assumption is consistent with the finding that people's confidence judgments are best calibrated for medium difficulty items and become overconfident as difficulty increases and underconfident as difficulty decreases (Suantak, Bolger, & Ferrell, 1996). Importantly, when people perform worse than chance (i.e., $p(C) < .5$), then $d' < 0$, indicating that the individual has a worse-than-chance discrimination ability. However, the individual's confidence in a decision is still based only on the subjective intensity because one cannot know whether one is correct or wrong in any particular trial. Because we assumed a fixed set of category boundaries, calibrated for medium difficulty items, this implies that for $d' < 0$, *higher* confidence implies a *lower* chance of being correct. This implication of the simulation setup is validated in the empirical results in this paper, where we show that the discrimination ability of people, as revealed by their confidence judgments, is indeed negative for wicked items where most people choose the wrong answer.

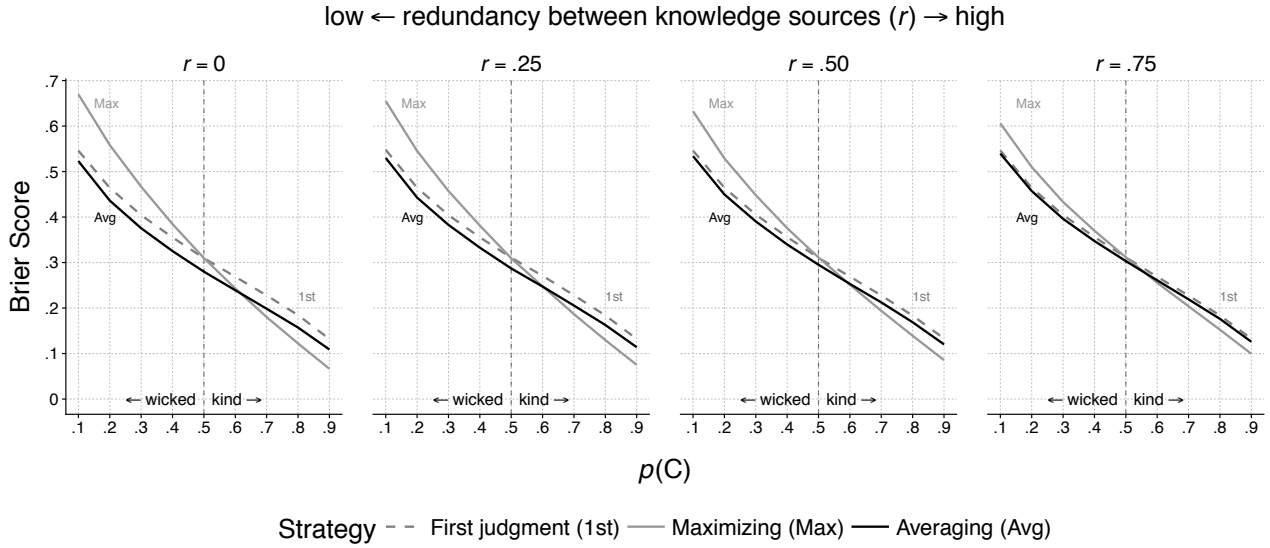


Figure 1. Overall accuracy of simulated confidence judgments as measured by the Brier score (y axis), where lower values indicate better quality. Panels (from left to right) correspond to increasingly more redundant knowledge sources underlying the two confidence judgments (correlation values r). The x axis shows the probability of being correct, where values of $p(C) > .5$ represent increasingly kinder items and values of $p(C) < .5$ increasingly more wicked items. Averaging outperformed first judgments, irrespective of the environment (more kind or more wicked items). Maximizing, in contrast, outperformed first confidence judgments only in kind environments (i.e. $p(C) > 0.5$), averaged judgments only for clearly kind environments (i.e. $p(C) > 0.6$). The effects of both aggregation strategies decreased as redundancy in knowledge sources increased.

confidence judgments strongly depended on the wickedness of the environment. Maximizing improved the Brier score in kind environments (i.e., $p(C) > .5$), for example, by .065 points for $r = 0$ and $p(C) = .9$, but impaired the Brier score in wicked environments (i.e., $p(C) < .5$), for example, by .09 points for $r = 0$ and $p(C) = .2$. Furthermore, maximizing outperformed averaging only once $p(C) > 0.6$ but not yet for $p(C) > 0.5$. As redundancy (r) increased, the sizes of these beneficial and harmful effects both decreased.

In real world environments, items typically differ in their probability $p(C)$ of being answered correctly. We therefore investigated the effects of averaging and maximizing in heterogeneous environments (for detailed results see Appendix , section A3). To summarize, the effects of averaging and maximizing depend simultaneously on the mean (μ) and variance of $p(C)$ of the environment. In general, as μ increased, the Brier score of all strategies improved. The effect of variance on the performance of confidence judgments depends on μ: In wicked environments ($\mu < .5$) increasing variance harmed the Brier score of all strategies, whereas in kind environments ($\mu > .5$) increasing variance improved the Brier score of first and averaged judgments, but continued to harm the Brier score of maximized judgments.

Some of these key results can also be ascertained analytically using a very general model that postulates for a particular item (1) the probability P that the *high*-confidence choice is correct, (2) the confidence C_H in this *high*-confidence choice, (3) the confidence C_L in the other, *low*-confidence choice, and (4) whether the high- and low-confidence choices are the same. Wicked items are characterized by $P < .5$ and thus imply that the high-confidence choice is more likely to be wrong than correct. Kind items, on the other hand, are

characterized by $P > .5$ and imply that the high-confidence choice is more likely to be correct than wrong. The main analytical insights are as follows (see Appendix A1 for details). First, for a wicked item (i.e., $P < .5$), averaging *always* has a better expected Brier score than maximizing, irrespective of whether or not the low-confidence choice is also wrong. Second, for a kind item (i.e., $P > .5$), the conditions are more complicated and depend on whether or not the low-confidence choice is also correct. When the high-confidence choice is very likely to be correct (i.e., $P \geq \frac{7}{8}$, that is, a very easy, kind item) but the low-confidence choice is wrong, the expected Brier score of maximizing is always better than that of averaging. In contrast, when both the low- and high-confidence choices are correct, there are no sufficient conditions that depend only on P for which maximizing always has a better expected Brier score than averaging. There are a series of conditions that specify for particular relationships between P, C_H, and C_L whether averaging or maximizing will have a better expected Brier score.

Apart from overall accuracy (in terms of, for example, the Brier score), confidence judgments can be evaluated along several dimensions of accuracy, including *calibration* (i.e., the extent to which subjective and objective probabilities match) and *resolution* (i.e., the extent to which confidence discriminates between correct and wrong decisions, irrespective of calibration). We assessed the resolution by calculating the DI' score:

$$DI' = \frac{slope}{\sqrt{scatter}},$$

which is the difference between mean confidence of correct vs. incorrect decisions (i.e., slope), standardized by the pooled SD of confidence judgments (i.e., scatter). [3] To investigate how calibration and resolution contribute to overall accuracy (in terms of the Brier score) and how they are influenced by the environment and the dependency among knowledge sources, we decomposed the Brier score using the covariance decomposition (Yates, 1990).

The results further validate the simulation setup (see Appendix A3). Here we highlight the most important set of findings. As expected by the design of the simulation, for kind items (i.e., $p(C) > .5$, Figure 2), confidence judgments discriminated between correct and wrong decisions (i.e., *positive resolution*). For wicked items (i.e., $p(C) < .5$), however, confidence judgments *wrongly* discriminated between correct and wrong decisions; that is, as items became more wicked, confidence increased for the wrong decision and decreased for correct decisions (i.e., *negative resolution*). This pattern of results is consistent with Koriat's consensuality principle (Koriat, 2012a): Confidence correlates with the size of the majority of people who favor one of the two possible answers (indexed by $p(C)$ in our simulation) and not with accuracy per se. By the very nature of maximizing, this implies that maximizing will improve resolution for kind items but *worsen* it for wicked items—a result we obtained. Averaging had an effect on resolution similar to that of maximizing, but it performed better on two other measures of the decomposition (bias and scatter) and therefore outperformed maximizing for wicked items.

[3] With $slope = \overline{conf}_{correct} - \overline{conf}_{wrong}$ and $scatter = \frac{n_{correct} var(conf_{correct}) + n_{wrong} var(conf_{wrong})}{n_{correct} + n_{wrong}}$.

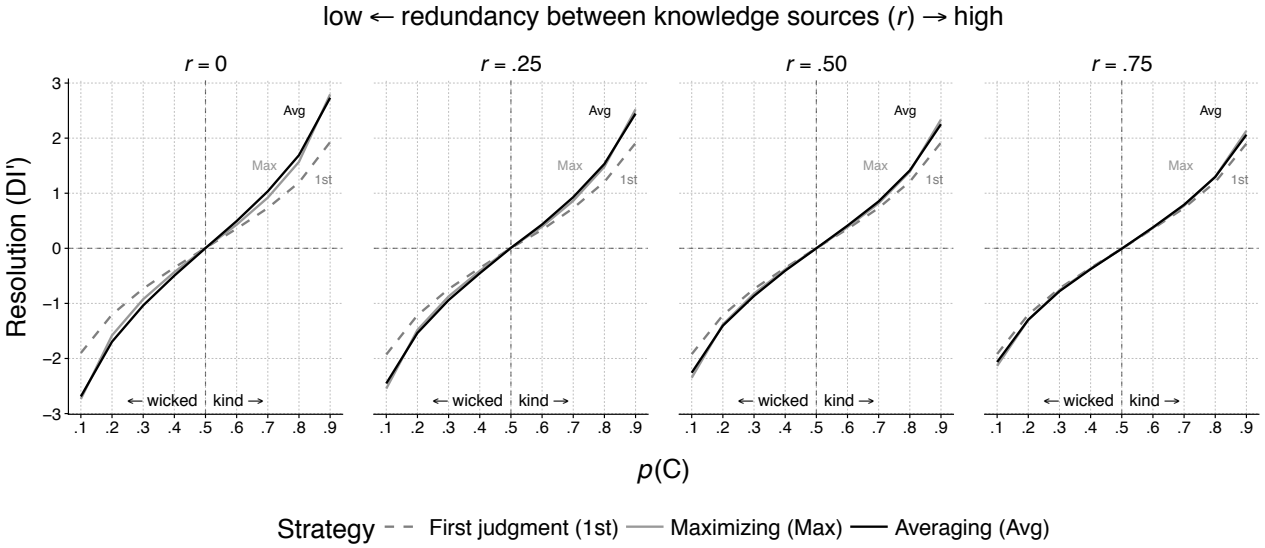


Figure 2. Resolution of simulated confidence judgments as measured by DI' (y axis). DI' quantifies the ability of confidence judgments to discriminate between correct and wrong decisions (i.e., difference between mean confidence of correct vs. incorrect decisions, standardized by the pooled SD of confidence judgments). Values above 0 indicate better discrimination; values below 0 indicate increasingly wrong discrimination, that is, confidence in the wrong decision is higher than in the correct decision. Panels (from left to right) correspond to increasingly more redundant knowledge sources underlying the two confidence judgments (correlation values r). The x axis shows the probability of being correct, where values of $p(C) > .5$ represent increasingly kinder items and values of $p(C) < .5$ represent increasingly more wicked items. Averaging and maximizing performed similarly: They outperformed first judgments for kind items but fell behind for wicked items.

Summary

Our simulation analysis, based on a signal detection framework of confidence (Ferrell & McGoey, 1980; Gu & Wallsten, 2001), investigated how the kindness versus wickedness of the environment (i.e., the degree to which people tend to choose the correct or wrong answer) and redundancy in knowledge sources used affect the performance of averaging and maximizing. The simulation study produced four major insights. First, averaging judgments resulted in improved overall accuracy (i.e., reduced Brier score) irrespective of the wickedness of the items. Second, for wicked items, maximizing judgments resulted in poorer accuracy than sticking to the first judgment but in better accuracy for kind items. These findings are further supported by our analytical analysis, showing that for wicked items, averaging necessarily *always* has a better expected Brier score than maximizing. Third, maximizing outperformed averaging only for items where $p(C) > 0.6$, but not yet for $p(C) > 0.5$. That is, a kind item is a necessary but not a sufficient condition for maximizing to outperform averaging. Finally, confidence correlated with how strongly the majority agreed on an answer, not with the correctness of the decision per se, and this partly explains why maximizing wicked items results in poorer overall accuracy (i.e., increased Brier score) compared to averaging wicked items.

What are the prescriptive recommendations that can be made on the basis of these results? Even when informed about the presence of wicked items, people have been found to lack the necessary insights to know whether an item is likely to be kind or wicked (Koriat, 2015, 2017). This means that relying on maximizing is

a bit of a gamble; yet, the risk in the gamble is attenuated by the fact that when $p(C) > .6$ maximizing does as well or better than averaging. In contrast, averaging one's first and second confidence judgments should always improve the overall accuracy of confidence judgments, even for wicked items, and therefore averaging can be used to one's benefit even though people cannot tell whether an item is kind or wicked.

However, as the simulation showed, all these effects were smaller the higher the redundancy among the knowledge sources underlying the two confidence judgments. Because actual confidence judgments within people are quite redundant (Ariely et al., 2000)—as we will show, the median correlation between two confidence judgments ranged between 0.5 and .0.85 across our empirical datasets—it could be that people's confidence judgments are so highly correlated that the differences between the strategies were not meaningful and thus largely irrelevant. Furthermore, it could also be that some assumptions of the simulation do not hold well enough for actual confidence judgments and therefore there remains the risk that the simulation analysis' insights might simply prove insufficient, and so, by extension, any recommendations based on them. When Ferrell and McGoey (1980) tested their signal detection model of confidence against empirical data, they noted that the empirical analyses corroborated many of the important qualitative patterns predicted by their model, but they also found several systematic differences. For example, their model was less able to model decisions about verbal assertions as compared to perceptual stimuli.

For all the above reasons, we investigated, using three empirical studies, how well the insights from our theoretical analysis generalize to individuals' actual confidence judgments as well as their practical relevance. On the basis of the results from our analysis, we investigate the following expected regularity: Always averaging an individual's two confidence judgments results in higher overall accuracy than either always maximizing confidence or always choosing the first confidence judgment. In the following, we reanalyze two published experiments and report on a new experiment we conducted.

The Performance of Averaging Versus Maximizing Confidence Judgments: Three Empirical Studies

To the best of our knowledge, there has hitherto been only one study that has investigated averaging confidence judgments *within* people (Ariely et al., 2000). That study reported only a small benefit of averaging on the quality of confidence judgments relative to averaging between people and attributed that to the higher redundancy in confidence judgments within relative to between participants. Similarly, there has so far been only one study that has investigated the effects of selecting the decision with the higher confidence judgment *within* a person (maximum-confidence-slating (MCS) technique; Koriat, 2012b). Koriat's MCS technique, however, is mute about the confidence one should place in the maximum-confidence decision. Koriat evaluated the accuracy of the maximum-confidence decisions (correct vs. wrong) but not that of the maximum-confidence judgments themselves (e.g., Brier score). Moreover, his analysis reported the accuracy of maximum-confidence decisions separately for kind and wicked items. For kind items, that is, where the majority of people chose the correct option, Koriat found a slightly higher percentage of correct answers (82%) for maximizing decisions

Case 1: Different decisions and different confidence judgments

		Brier Score	
1st judgment	90% Bulgaria (equivalent to 10% Romania)	$(.9_{Bulgaria} - 1)^2 = .01$	
2nd judgment	70% Romania (equivalent to 30% Bulgaria)	$.7_{Romania} - 0)^2 = .49$	
Averaging	$\frac{90_{Bulgaria}+30_{Bulgaria}}{2} = 60_{Bulgaria}$	$(.6_{Bulgaria} - 1)^2 = .16$	Choice of reference class is irrelevant for the Brier score
	or		
	$\frac{10_{Romania}+70_{Romania}}{2} = 40_{Romania}$	$(.4_{Romania} - 0)^2 = .16$	
Maximizing	$90_{Bulgaria}$	$(.9_{Bulgaria} - 1)^2 = .01$	

Case 2: Different decisions same confidence judgments

		Brier Score	
1st judgment	70% Bulgaria (equivalent to 30% Romania)	$(.7_{Bulgaria} - 1)^2 = .09$	
2nd judgment	70% Romania (equivalent to 30% Bulgaria)	$.7_{Romania} - 0)^2 = .49$	
Averaging	$\frac{70_{Bulgaria}+30_{Bulgaria}}{2} = 50_{Bulgaria}$	$(.5_{Bulgaria} - 1)^2 = .25$	Choice of reference class is irrelevant for the Brier score
	or		
	$\frac{30_{Romania}+70_{Romania}}{2} = 50_{Romania}$	$(.5_{Romania} - 0)^2 = .25$	
Maximizing	$70_{Bulgaria}$	$(.7_{Bulgaria} - 1)^2 = .09$	Choice of reference class is relevant for the Brier score
	or		
	$70_{Romania}$	$(.7_{Romania} - 0)^2 = .49$	

Table 1. Applying averaging and maximizing when decisions or confidence judgments differ. Different decisions, but equal confidence judgments occurred in Study 1 (Ariely et al., 2000) in 1.3% of the trials, in Study 2 (Koriat, 2012b) in 1.4% of the trials and in Study 3 (New Experiment) in 0% of the trials.

compared to the typical performance of first and second judgments (81%). For wicked items, that is, where the majority of people choose the wrong option, the percentage of correct answers dropped to 24% when maximizing, whereas the typical performance of first and second judgments was now slightly higher at 25%.

In contrast to Koriat (2012b), we investigated whether maximizing can increase the accuracy of confidence judgments and how useful this strategy is without knowing the kindness versus wickedness of an item. Assuming that individuals do not know beforehand what type of item they face (Koriat, 2015, 2017), we investigated whether it is possible to improve the quality of confidence judgments by always applying either averaging or maximizing. To this end, we analyzed averaging and maximizing in two datasets, where participants indicated their confidence about which of two U.S. cities has a larger population (Ariely et al., 2000) or about which of two geometric figures was longer or larger, respectively (Koriat, 2012b). Table 1 illustrates the implementation of averaging and maximizing, given that people may, when asked again, not only indicate a different level of confidence, but also choose the other answer.

Furthermore, we conducted a study to test whether *dialectical* bootstrapping (Herzog & Hertwig, 2009, 2013, 2014a), a framework aiming to reduce redundancy in an individual's estimates by using suitable elicitation

techniques, could reduce redundancy in confidence judgments and as a result enhance the effects of averaging. Herzog and Hertwig (2009) first tested the dialectical bootstrapping approach in a quantitative estimation task using the consider-the-opposite technique (adapted from Lord, Lepper, & Preston, 1984). More precisely, in their experiment, participants were told to assume that their first estimate was off the mark, to think about reasons why that could be, and to produce a second, "dialectical" estimate. They found that averaging dialectical estimates led to larger gains in accuracy than simply averaging repeated estimates. In our new experiment, we tested whether applying the dialectical bootstrapping approach (using the consider-the-opposite technique) can also reduce redundancy in confidence judgments about general knowledge questions (e.g., "Who was the tutor of Alexander the Great first? (a) Aristotle or (b) Plato"), and whether, as a consequence, averaging dialectical judgments can improve the overall accuracy further, compared to averaging merely repeated judgments. To the best of our knowledge, this is the first test of dialectical bootstrapping in the service of boosting the wisdom of the inner crowd in the context of confidence judgments. We made no predictions about how the consider-the-opposite technique would influence the accuracy of maximizing. All data and scripts to reproduce the empirical analyses can be found at: https://osf.io/b3f6d/?view_only=22b543c3ab3f4943af67b5c4842127d5

Methods

Study 1 (Ariely et al., 2000). The first dataset comes from a study by Ariely et al. (2000, referred to as Study 3 (New Experiment) in their article) involving representative questions about the population sizes of the 50 largest cities in the United States in 1992. Sixty-four students of the University of North Carolina, Chapel Hill participated and were paid a minimum of $4 plus a bonus that depended on their performance. The questions about the relative sizes of two cities were presented as either single true-or-false statements (TF) or complementary pairs of statements (PC) written above each other, where one was the opposite of the other. Participants indicated their belief in the statements with confidence judgments ranging from 0% to 100%, without providing a decision (true vs. not true), and later, in the same session, they assessed the same statements again. For a more detailed description refer to Ariely et al. (2000). We made no predictions about whether or how the results would differ depending on the response format (TF vs. PC).

Study 2 (Koriat, 2012b). The second dataset comes from Koriat (2012b, referred to as Study 5 in his article). Fifty University of Haifa psychology undergraduates (43 females, 7 males) were asked to compare the areas of geometric shapes and the lengths of irregular lines. The shapes task deliberately included more wicked items (40%) than the lines task (20%). Participants first chose the larger object and then assigned their confidence in their decision on a half-range probability scale (50–100%). The study consisted of two sessions with a 1-week interval between them. For a more detailed description see Koriat (2012b). The higher number of wicked items in the shapes task should put the maximizing strategy at a higher risk to do more harm than good compared to the lines task, which featured fewer wicked items. Beyond that we made no predictions about whether the results differ depending on the shapes or line task.

Study 3 (New Experiment).

Participants. The data collection occurred at a previous institution (University of Basel, Switzerland). As this experiment was a non-clinical study and did not involve any patients, it did not classify as requiring in-depth evaluation and approval by a cantonal review board according to Swiss federal law. A total of 309 (160 female, 149 male) U.S. participants were recruited via Amazon Mechanical Turk for an approximately 45-min survey and were reimbursed with a flat fee of \$2.[4] Forty-eight participants did not pass the instructional manipulation check (i.e., a question testing their attention) and were thus excluded from further analyses. The experiment deliberately did not force participants to only enter confidence judgments between 50% and 100%, to thus be able to monitor their attention to the task. When participants gave an answer outside of the permissible range, we treated this trial as missing. Five participants were excluded because they gave more than three answers outside this range. Furthermore, 25, 5 and 1 participants gave 1, 2 and 3 confidence judgments, respectively, outside the range.

Materials and procedure. The material was taken from Gigerenzer et al. (1991) and included 50 general knowledge questions about history, nature, geography, and literature (e.g., "Sofia is the capital of: (a) Romania or (b) Bulgaria?). This question set deliberately included wicked items. In a pretest we created two comparable subsets of 25 items each, which were matched by proportion correct, bias, and Brier scores. We used one of these subsets in the main study here (see Appendix A2, Table A1). Participants provided their decision first and then assigned their confidence on a half-range probability scale (50%–100%). The experiment was split into two sessions. In the first session, participants answered the 25 questions. In the second session, participants were allocated either to the *dialectical condition* or to the *reliability (control) condition* and responded to the same questions again. After answering all 25 questions for a second time, participants were directed to the online form of the new Berlin numeracy test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). We administered this measure for exploratory purposes and have not yet analyzed its data.

In the dialectical condition (n = 119), participants were asked to generate dialectical decisions and corresponding confidence judgments while we showed them their first decision and confidence judgment (Herzog & Hertwig, 2009, 2014a). The consider-the-opposite instructions (adapted from Lord et al., 1984) read:

> First, assume that your first answer and confidence judgment were off the mark. Second, think about a few reasons why that could be. Which assumptions and considerations could have been wrong? Third, what do these new considerations imply? Does your answer make sense? Was the first confidence judgment rather too high or too low? Fourth, based on this new perspective, give a new answer and indicate your confidence in it. Please feel free to totally change your mind.

[4] On the basis of the medium effect of the dialectical instruction on the accuracy of quantitative estimates observed in Herzog and Hertwig (2009, p. 234; Cohen's d = 0.53), we considered a small to medium effect of the dialectical instruction on the accuracy of confidence judgments as plausible a priori. We aimed for a sizeable sample size of n = 150 per condition and recruited a few more participants in the anticipation that we would need to exclude a few who did not follow instructions.

In the reliability condition (n = 137), participants were not shown their first responses and were instructed to answer the questions as if they were seeing them for the first time.

Statistical analyses. After calculating accuracy measures for first, second, averaged, and maximized confidence judgments, we conducted a Bayesian parameter estimation analysis (Kruschke, 2013) of the differences between accuracy measures of first minus averaged and first minus maximized judgments. For the majority of measures, first and second judgments did not differ systematically throughout the three studies; the sole exception was that in the TF condition in Study 1 (Ariely et al., 2000) second judgments had a better Brier score. We therefore report differences between first and averaged and first and maximized confidence judgments. Comparing second to averaged and maximized confidence judgments qualitatively yielded largely the same results. We conducted our analyses in the statistical computing software R and used the default priors from the BEST package (Kruschke & Meredith, 2015). The resulting posterior distributions of the parameters illustrate the credibility of the values given the data. We summarize the posterior distributions by reporting medians as point estimates and 95% highest density intervals (HDIs) as uncertainty intervals. A 95% HDI expresses the uncertainty around the estimate and states in which interval the true value is likely to fall with a 95% probability (according to the model). When displaying effect sizes in figures, we highlight a "region of practical equivalence," for which Cohen's d's effect size is conventionally considered to be small (from −0.1 to +0.1).

Results

Environments. Figure 3 shows the distribution of proportion correct across items in Studies 1–3. Study 1 (Ariely et al., 2000) and Study 2 (Koriat, 2012b) (Figure 3, panels A and B) contained more wicked items than Study 3 (New Experiment) (Figure 3, panel C), thereby putting the maximizing strategy at risk of doing more harm than good.

Confidence—kindness/wickedness relationship. Figure 4 depicts the relationship between the kindness/wickedness of an item and the ability of participants' first confidence judgments to discriminate between correct and wrong answers (as measured by DI'). Consistent with the simulation study, the more strongly the majority agreed on the correct answer, the more clearly confidence distinguished between correct and wrong answers. Notably, the more strongly the majority agreed on the wrong answer, the more clearly confidence distinguished, albeit in a reversed fashion, between correct and wrong answers (i.e., as proportion correct per item fell below .5, discrimination became negative, that is, $DI' < 0$).

Redundancy in knowledge sources: Correlation between two confidence judgments within individuals. Figure 5 summarizes the distribution of Spearman correlations between first and second confidence judgments within participants across Studies 1–3 (median correlations ranged between .5 and .85). In Study 3 (New Experiment), the median correlation in the dialectical condition was lower ($r_{dialectical}$ = .77) than in

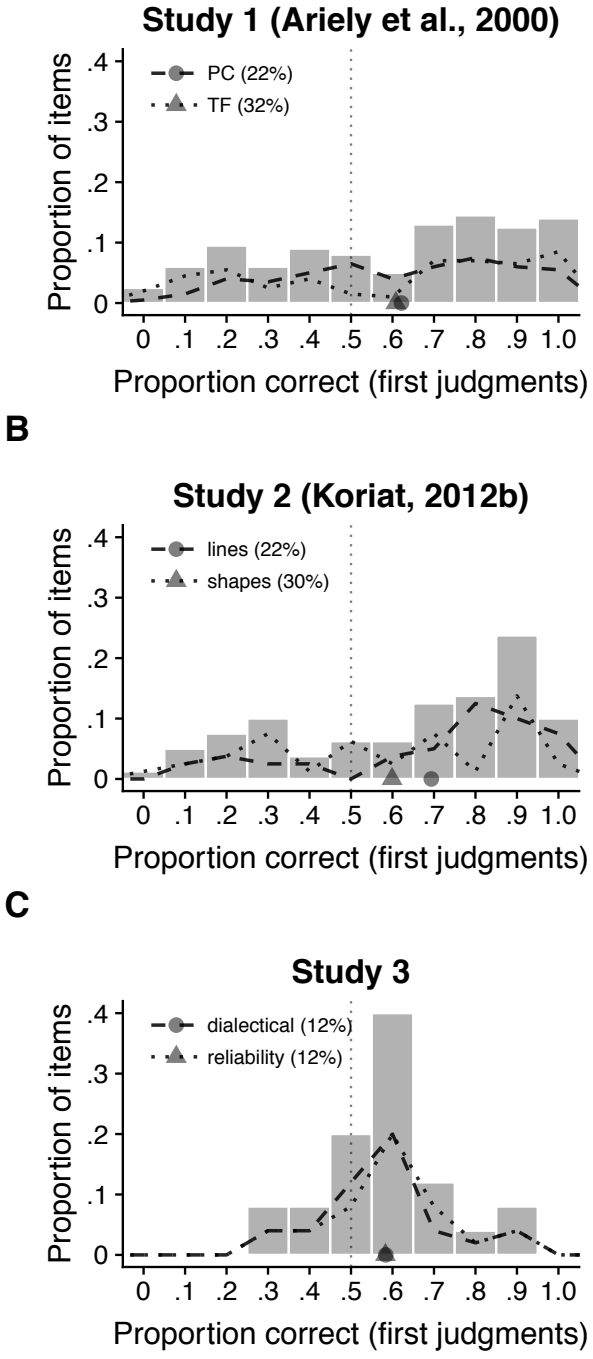


Figure 3. Histograms of proportion correct of items (based on first judgments), separately for each study. Dashed and dotted lines show the distributions per condition (A, C) or task (B). Circles and triangles on the bottom of each panel indicate median proportion correct across items per condition (A, C) or task (B). Legends report percentages of clearly wicked items (i.e., $p(C) < .4$) per condition (A, C) or task (B). Study 1 (Ariely et al., 2000) and Study 2 (Koriat, 2012b) contained more clearly wicked items than Study 3 (New Experiment). PC = pairwise comparison condition; TF = true-or-false condition.

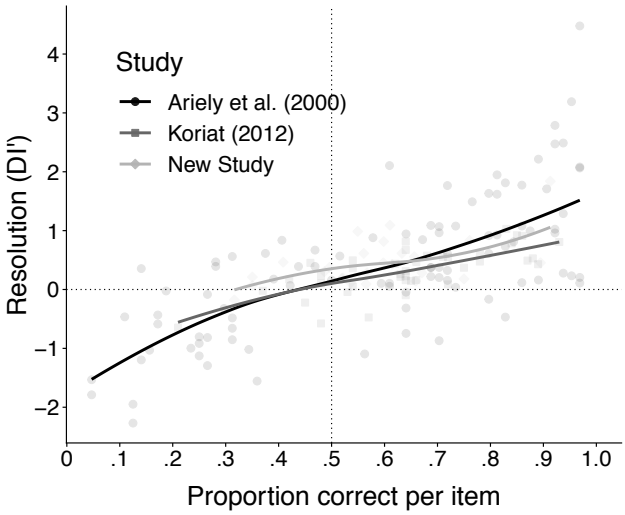


Figure 4. Confidence's ability to distinguish between correct and wrong first answers per item (DI' per item; y axis) as a function of proportion correct of that item (x axis). Results are shown separately for each study (pooled across the two tasks in Study 1 (Ariely et al., 2000) and Study 2 (Koriat, 2012b)). Circles, triangles and crosses indicate items per study, and smoothed lines show for each study a robust local polynomial regression (LOESS) fit. DI' quantifies the ability of confidence judgments to discriminate between correct and wrong decisions (i.e., difference between mean confidence of correct vs. incorrect decisions, standardized by the pooled SD of confidence judgments). Values above 0 indicate better discrimination; values below 0 indicate increasingly wrong discrimination, that is, confidence in the wrong decision is higher than in the correct decision. As items become more wicked, confidence increases for wrong decisions and decreases for correct decisions (i.e., *negative resolution*).

the reliability condition ($r_{reliability}$ = .85; Cohen's $d_{reliability-dialectical}$ = 0.7, 95% HDI [0.38, 1.03]). This suggests that the consider-the-opposite technique in the dialectical bootstrapping condition successfully reduced redundancy in participants' confidence judgments.

Overall accuracy of confidence judgments. To evaluate the effects of averaging and maximizing, we compared averaged and maximized confidence judgments against first judgments. On the basis of the results from our simulation analysis, we predicted that consistently averaging participants' confidence judgments would result in a higher overall accuracy than consistently maximizing their judgments.

Averaging versus first confidence judgments. Averaging consistently led to improved Brier scores throughout the three studies (Figure 6 and Table 2), even when median correlations between two confidence judgments were relatively high (e.g., Study 1: $r_{reliability}$ = .85), as well as when environments contained a substantial number of wicked items (e.g., Study 1 and Study 2).

Maximizing versus first confidence judgments. With one exception, maximizing consistently harmed Brier scores compared to first, initial confidence judgments throughout the three studies (Figure 6 and Table 2). Only in the lines task in Study 2 was the effect size not reliably different from a zero effect (d_{lines} = −0.015, 95% HDI [-0.37, 0.37]). Drawing on the insights from the simulation analysis, we suggest that the overall negative effect of maximizing can be partly explained by the respective number of wicked items in two tasks. In line with the relatively large number of clearly wicked items (i.e., $p(C) < 0.4$) in Study 1 (Ariely et al., 2000; Figure 3, panel A: 22% in the PC and 32% in the TF condition) and Study 2 (Koriat, 2012b; Figure 3, panel

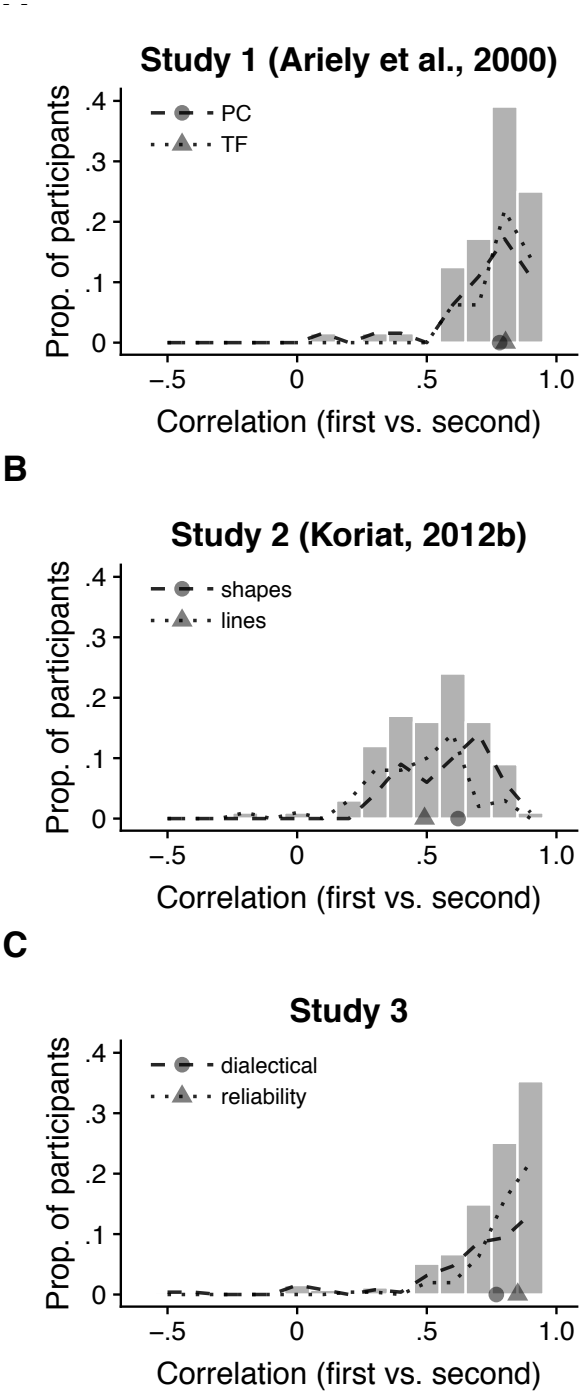


Figure 5. Histogram of Spearman correlations between first and second confidence judgments separately for study. Dashed and dotted lines show distributions per condition (A, C) or task (B). Circles and triangles on the bottom of each panel indicate median correlations per condition or task. Study 1 and Study 3 were run in one session, whereas Study 2 elicited repeated judgments after a 1-week interval, which could possibly explain the lower correlations compared to Study 1 and Study 3. PC = pairwise comparison condition; TF = true-or-false condition.

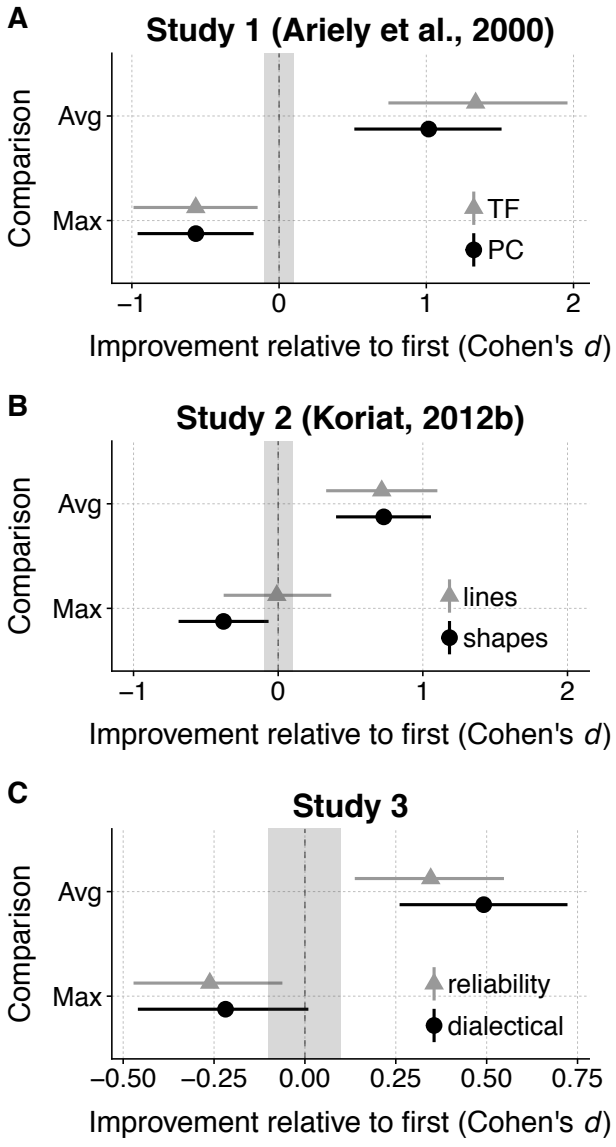


Figure 6. Effects of averaging (Avg) and maximizing (Max) on the overall accuracy of confidence judgments relative to first judgments. The x axis shows the improvement in Brier scores expressed as Cohen's d effect sizes; symbols show the median value and the ranges show the 95% highest density interval of the posterior distribution). Bars to the right of zero imply improved scores and bars to the left of zero imply harmed scores. The shaded region ranging between −0.1 and +0.1 highlights the region of practical equivalence, for which Cohen's d effect size is conventionally considered to be small (from −0.1 to +0.1). Averaging confidence judgments consistently and reliably outperformed the quality of first judgments throughout the three studies. Maximizing, in contrast, tended to harm and never improved the quality compared to first confidence judgments.

Table 2. Cohen's d Effect Sizes for Differences in Brier Scores Between First Versus Averaged and First Versus Maximized Confidence Judgments

Study	Condition	Cohen's d	95% HDI
First judgments vs. averaging			
Ariely et al. (2000)	PC	1.003	[0.511, 1.511]
	TF	1.317	[0.743, 1.959]
Koriat (2012b)	Shapes	0.728	[0.400, 1.056]
	Lines	0.707	[0.332, 1.100]
New study	Dialectical	0.490	[0.260, 0.722]
	Reliability	0.345	[0.137, 0.548]
First judgments vs. maximizing			
Ariely et al. (2000)	PC	-0.565	[-0.961, -0.173]
	TF	-0.560	[-0.989, -0.146]
Koriat (2012b)	Shapes	-0.377	[-0.689, -0.066]
	Lines	-0.015	[-0.377, 0.367]
New study	Dialectical	-0.216	[-0.460, 0.010]
	Reliability	-0.261	[-0.472, -0.062]

Note. PC = pairwise comparison condition; TF = true-or-false condition; Cohen's d = median value of the posterior distribution; 95% HDI = 95% highest density interval of the posterior distribution.

B: 30% in the shapes and 22% in the lines task), maximizing's harmful effect on the Brier score is large (e.g., $d_{PC} = -0.56$, 95% HDI [-0.96, -0.17]) or medium ($d_{shapes} = -0.38$, 95% HDI [-0.68, -0.06]), respectively. In contrast, Study 3 (New Experiment) contained relatively few clearly wicked items (Figure 3, panel C: 12% in both, the dialectical and reliability condition) and maximizing's harmful effect is small (e.g., $d_{reliability} = -0.26$, 95% HDI [-0.47, -0.06]).

Averaging dialectical versus reliability judgments. On the basis of the results of the simulation analysis, we expected that the effects of averaging would be moderated by the size of the correlation between first and second confidence judgments. In Study 3 (New Experiment), we investigated whether dialectical bootstrapping (Herzog & Hertwig, 2009, 2014b) successfully reduces the redundancy (i.e., correlation) in confidence judgments and whether, consequently, averaging first and dialectical judgments can further improve the overall accuracy compared to averaging first and merely repeated confidence judgments. As already reported above, the median correlation between participants' confidence judgments was lower in the dialectical bootstrapping condition ($r_{dialectical} = .77$) than in the reliability condition ($r_{reliability} = .85$). Consistent with our prediction, there is some evidence that averaging dialectical judgments may have enhanced the Brier score more than merely averaging reliability judgments ($d = 0.28$, 95% HDI [-0.02, 0.59]).

Decomposition of overall accuracy. Finally, to understand how averaging and maximizing contribute to the changes in overall accuracy, we conducted a Brier score decomposition (Yates, 1990, using the covariance decomposition), which yields estimates of calibration and resolution, as well as estimates for bias (over- vs.

underconfidence) and scatter (random error). Our analysis showed that gains from averaging were mainly driven by reduced bias, whereas losses from maximizing primarily resulted from increased bias (see Appendix A4 for detailed results).

General Discussion

Can the inner crowd be harnessed to boost accuracy of confidence judgments? We undertook the first comprehensive analysis of when and how two competing ways of harnessing the wisdom of the inner crowd (Herzog & Hertwig, 2014a)—maximizing or averaging individual's multiple and possibly conflicting confidence judgments—improves the accuracy of people's final confidence in their decision. We find that an individual can enhance the accuracy of her final confidence judgment by averaging her two confidence judgments (Ariely et al., 2000). In contrast, maximizing, that is, using the highest confidence judgment (Koriat, 2012b, adapted from the MCS technique) proves risky: It performs better than averaging for clearly kind items, but worse otherwise. Next, we first review implications from our simulation and empirical analysis for the effects of maximizing and averaging. We then discuss the limitations of our simulation analysis and the boundary conditions for aggregating ever more judgments from the same person. Finally, we conclude by relating our research to the phenomenon of the wisdom of crowds and the literature on other strategies to improve confidence judgments.

Boundary Conditions for Averaging and Maximizing Confidence Judgments

An individual evaluates the same item on two different occasions, and each time produces a confidence judgment. What should the individual do to improve the accuracy of these confidence judgments? One strategy is to average them. Another one is to select the highest confidence judgment. We investigated the performance of both strategies analytically and by simulating different items (i.e., questions) ranging from those for which most people would make correct decisions ("kind" items) to those for which *most* people would make wrong decisions ("wicked" items). Our analytical and simulation results suggest that if an individual averages the confidence judgments, then their overall accuracy would be improved, even for wicked items. Maximizing, in contrast, proves risky. It outperforms averaging only for clearly kind items ($p(C) > .6$). In light of the fact that people appear to lack the necessary skills to assess the kindness vs. wickedness of a question in advance (Koriat, 2015, 2017), our analysis suggests that averaging—due to its robustness—is the strategy that the individual should apply to best exploit her conflicting confidence judgments.

One possible limitation of our analysis is the assumption that first and second confidence judgments do not differ in their discrimination ability. Since we mostly did not find that first and second confidence judgments differed in the empirical datasets, this assumption seems realistic. Future research could nevertheless extend the predictions of the simulation to investigate the influence of differing discrimination abilities and calibration of first and repeated confidence judgments on the performance of averaging and maximizing.

Since actual repeated confidence judgments from the same person are substantially correlated (Ariely et al., 2000), we reanalyzed datasets from two previously published studies and conducted one new study to

investigate whether the results from the simulation analysis generalize to empirical confidence judgments. The median correlations in our empirical datasets ranged between .5 and .85 (see also Figure 5). Consistent with the simulation analysis, we found that averaging two confidence judgments from the same person improved overall accuracy (i.e., Brier score), whereas maximizing among a person's confidence judgments harmed overall accuracy, even in environments with relatively few wicked items (i.e., Study 3; see Figure 3).

We considered settings in which a person produced two confidence judgments. At least in theory, it is conceivable that a person produces even more confidence judgments. Would averaging or maximizing them further increase accuracy? Averaging more confidence judgments generated by the same person would unlikely result in notably higher averaging gains, because error redundancy in a person's judgments places an upper limit on the effect of averaging (Rauhut & Lorenz, 2011; Van Dolder & van den Assem, 2018). In contrast, maximizing over an increasingly larger set of confidence judgments from the same person is likely to further amplify the effects we found for maximizing because making more and more judgments renders it increasingly more likely that an even higher confidence judgment will be generated.

The Wisdom of Crowds: Averaging and Maximizing Confidence Judgments Across Individuals

The insights from our analysis apply to judgment aggregation strategies both within and *between* individuals because the simulated confidence judgments can be viewed as stemming from the same person or two different people. Because judgments from different people are less redundant than the same person's judgments (Herzog & Hertwig, 2014a), our analysis predicts stronger effects when judgments are aggregated between non-interacting people (see the panels in Figure 1 with lower knowledge redundancy). Furthermore, the returns from averaging more people will diminish more slowly (see also Rauhut & Lorenz, 2011; Van Dolder & van den Assem, 2018) and the effects of maximizing across ever more people should be even more pronounced compared to combining ever more confidence judgments from the same person (as discussed in the previous subsection).

Alternative Methods for Improving Accuracy of Confidence Judgments

Averaging and maximizing represent two of the many strategies that have been proposed for improving the accuracy of confidence judgments. For example, recalibrating individual confidence judgments when aggregating forecasts of several individuals has been shown to improve forecast accuracy by 26% (Turner, Steyvers, Merkle, Budescu, & Wallsten, 2014). Furthermore, Baron et al. (2014) show that averaged confidence judgments should be extremized because of at least two processes, which render individual confidence judgments too regressive: (i) random error can only be distributed asymmetrically towards 0.5 the closer one's internal, latent confidence is to one of the end points of the probability scale; and (ii) awareness of one's incomplete knowledge may lead individuals to preemptively regress their confidence judgments towards 0.5. This latter

process can be appropriate when the goal is to increase individual accuracy, but will typically result in too conservative confidence judgments when the goal is aggregate them. [5]

Other strategies aim to improve the quality of confidence judgments by trying to reduce overconfidence, for example, by urging people to consider evidence inconsistent with their current beliefs (Koriat et al., 1980) or alternative outcomes and explanations (Hirt & Markman, 1995). Yet, these techniques are typically evaluated solely in the context of overconfidence (Arkes, 2001). Our work shows that a much richer analysis would consider not only the effects of these different strategies on over- vs. underconfidence, but on the overall Brier score as well as its different components and how different statistical environments impact the effectiveness of these strategies.

Our own analysis has of course limitations. One is that our signal detection model of confidence judgments is a static model that can be understood as people basing their confidence judgments on a fixed sample of evidence about whether or not their decision is likely to be correct. However, recent work has begun to show that confidence is based on a dynamic process where sequential samples of evidence are accumulated over time (Pleskac & Busemeyer, 2010; Yu et al., 2015). From this perspective, differences between averaging and maximizing depend in part on how the second confidence judgment is being generated. In our current datasets participants provided two confidence judgments that were either spaced out within the same (Study 1 and Study 3) or a different (Study 2) experimental session. Thus, both confidence judgments were the result of two separate evidence accumulation processes, and assuming all else held constant, our results suggest averaging being superior to maximizing across kind and wicked environments. However, now consider a context in which individuals are asked to make two sequential confidence judgments in response to the same question and in close temporal proximity. According to Pleskac and Busemeyer's (2010) model, individuals continue to accumulate evidence even after they have made an initial response. Thus, the second judgment is likely to be based on even more accumulated evidence than the first judgment. Now how would one best harness the wisdom of the inner crowd taking this dynamic perspective into account? This is an interesting question that deserves more theory and experimentation. Our tentative answer is that it depends on the item. If the item is kind, then the second confidence judgments will eventually yield a better resolution than the first judgments. As a consequence, selecting the second confidence judgments should be a superior strategy to averaging both confidence judgments. In other words, for kind items, from a dynamic perspective, when confidence judgments are generated in close temporal proximity one should not average or maximize but should categorically select

[5] When aggregating judgments within the same individual, we would likewise expect both the end-of-scale and the confidence-regression effects to occur. However, the overall regression of averaged confidence towards 0.5 (and thus the need for extremizing) should be less pronounced than in the case of different people. Because an individual's repeated judgments are more redundant than those of different individuals, regressing one's confidence towards 0.5 will underappreciate the information contained in a within-person average less as compared to the a between-person average . In contrast, the implications of the end-of-scale effect for extremizing should be the same, irrespective of whether averaging happens within or across individuals. Concerning averaging within people, any factor that increases aggregation gains (e.g., less redundancy in knowledge sources used at both occasions; i.e., smaller r in our simulation) would change the degree to which people's tendency to regress their confidence judgments will underappreciate the information contained in the average. This would then call for more extremizing, but likely still less than for averaging the same number of confidence judgments from difference people since those judgments will typically be still less redundant than those of one person. Furthermore, the moderation of these effects by the distribution of kind and wicked items should hold equally for maximizing as well as extremizing. The kinder the items, the more beneficial it is to extremize, and the more wicked the items, the more harmful it is to extremize.

the second judgment. For wicked items, in contrast, one should not select the second but the first confidence judgment. This is because for wicked items further evidence accumulation is likely to lead the decision maker further astray.

However, as people seem to lack the necessary skills to assess the kindness versus wickedness of an item in advance (Koriat, 2015, 2017), always choosing the first or the second judgment is again a risky strategy. In the absence of reliable knowledge on the type of item, averaging should perform better and be the preferred strategy—again. These ideas illustrate the importance of considering not only the environment, but also the cognitive processes in developing and prescribing methods for improving the accuracy of confidence judgments.

Conclusion

The wisdom of the inner crowd refers to the idea that individuals can harness their own multiple, perhaps even conflicting judgments pertaining to the same problem to improve the quality of their judgments (Herzog & Hertwig, 2014a). The study of ecological rationality (Todd, Gigerenzer, & ABC Research Group, 2012) involves asking the questions: Given a cognitive strategy, in what environments does it succeed? And given an environment, what cognitive strategies succeed in it? We asked these questions about the maximizing and averaging strategy applied to multiple confidence judgments of the same person. Our theoretical and empirical results suggest that averaging should be the preferred strategy to harness the wisdom of one's inner crowd. The reason is that the robust averaging strategy, relative to the more fickle maximizing strategy, can boost accuracy of confidence judgments while requiring less knowledge about the kindness and wickedness of the items the decision maker faces.

References

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., ... Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*, 130–147. doi: 10.1037/1076-898X.6.2.130

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (p. 495–515). Norwell, MA: Kluwer Academic.

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. doi: 10.1016/j.concog.2014.02.002

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 929–945. doi: 10.1037/0096-1523.24.3.929

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*, 133–145. doi: 10.1287/deca.2014.0293

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5 Suppl), S2–S23. doi: 10.1016/j.amjmed.2008.01.001

Betts, R. K. (1978). Analysis, war, and decision: Why intelligence failures are inevitable. *World Politics*, *31*, 61–89. doi: 10.2307/2009967

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, *78*(1), 1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The berlin numeracy test. *Judgment and Decision Making*, *7*, 25–47. doi: 10.1037/t45862-000

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*(6), 959–988. doi: 10.1037/0033-2909.130.6.959

Dougherty, M. R. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General*, *130*(4), 579–599. doi: 10.1037/0096-3445.130.4.579

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519–527. doi: 10.1037/0033-295X.101.3.519

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance*, *26*(1), 32–53. doi: 10.1016/0030-5073(80)90045-8

Garrett, H. E. (1922). *A study of the relation of accuracy to speed* (Vol. 56). Columbia university.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. doi: 10.1037/0033-295X.98.4.506

Griffin, D. W., & Brenner, L. A. (2004). Perspectives on probability judgment calibration. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 177–199). Oxford, England: Blackwell.

Gu, H., & Wallsten, T. S. (2001). On setting response criteria for calibrated subjective probability estimates. *Journal of Mathematical Psychology*, *45*(4), 551–563. doi: 10.1006/jmps.2000.1337

Hertwig, R. (2012). Tapping into the wisdom of the crowd—with confidence. *Science*, *336*(6079), 303–304. doi: 10.1126/science.1221403

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237. doi: 10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2013). The crowd within and the benefits of dialectical bootstrapping: A reply to white and antonakis (2013). *Psychological Science*, *24*(1), 117–119. doi: 10.1177/0956797612457399

Herzog, S. M., & Hertwig, R. (2014a). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506. doi: 10.1016/j.tics.2014.06.009

Herzog, S. M., & Hertwig, R. (2014b). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 218–232. doi: 10.1037/a0034054

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology*, *69*(6), 1069–1086. doi: 10.1037/0022-3514.69.6.1069

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 1068–1074. doi: 10.1037/a0019694

Johnson, D. D. P., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, *477*(7364), 317–320. doi: 10.1038/nature10384

Johnson, D. M. (1939). *Confidence and speed in the two-category judgement* (Vol. 34) (No. 241). Columbia university.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review*, *107*(2), 384–396. doi: 10.1037/0033-295X.107.2.384

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*(2), 123–141. doi: 10.1016/0010-0277(82)90022-1

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273. doi: 10.1016/0001-6918(91)90036-Y

Koriat, A. (1976). Another look at the relationship between phonetic symbolism and the feeling of knowing. *Memory & Cognition*, *4*(3), 244–248. doi: 10.3758/BF03213170

Koriat, A. (2008). Subjective confidence in one's answers: the consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 945–959. doi: 10.1037/0278-7393.34.4.945

Koriat, A. (2012a). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. doi: 10.1037/a0025648

Koriat, A. (2012b). When are two heads better than one and why? *Science*, *336*(6079), 360–362. doi: 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. doi: 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. doi: 10.1002/bdm.2024

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, *6*(2), 107–118. doi: 10.1037/0278-7393.6.2.107

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi: 10.1037/a0029146

Kruschke, J. K., & Meredith, M. (2015). Best: Bayesian estimation supersedes the t-test cran. *R-project.*

org/package= BEST (R package version 0.4.0.). Retrieved from https://CRAN.R-project.org/package=BEST

Kurvers, R. H. J. M., 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 of the United States of America*, *113*(31), 8777–8782. doi: 0.1073/pnas.1601827113

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers in social psychology: Social judgment and decision making* (p. 227—242). New York, NY: Psychology Press.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. doi: 10.1287/mnsc.1050.0459

Laughlin, P. R. (1980). Social combination processes of cooperative problem-solving groups on verbal intellective tasks. In M. Fischbein (Ed.), (Vol. 1, pp. 127–155). Hillsdale, NJ: Erlbaum.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189. doi: 10.1016/0022-1031(86)90022-3

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*(6), 969–998. doi: 10.1111/j.1551-6709.2009.01045.x

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231–1243. doi: 10.1037/0022-3514.47.6.1231

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(30), 10984–10989. doi: 10.1073/pnas.1406138111

McClelland, A. G., & Bolger, F. (1994). The calibration of subjective probability: Theories and models 1980–94. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Oxford, England: John Wiley & Sons.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. doi: 10.1177/0956797614524255

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. In G. Keren & G. Wu (Eds.), *The wiley blackwell handbook of judgment and decision making* (pp. 82–209). Chichester, UK: Wiley.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, *26*, 125–141. doi: 10.1016/S0079-7421(08)60053-5

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. doi: 10.1037/a0019737

Rauhut, H., & Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology*, *55*(2), 191–197. doi: 10.1016/j.jmp.2010.10.002

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*(2), 344–356. doi: 10.1016/j.ijforecast.2013.09.009

Sniezek, J. A., Paese, P. W., & Switzer III, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes*, *46*(2), 264–282. doi: 10.1016/0749-5978(90)90032-5

Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (2014). Measuring the crowd within again: a pre-registered replication study. *Frontiers in Psychology*, *5*, 786–794. doi: 10.3389/fpsyg.2014.00786

Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, *10*(7), 327–334. doi: 10.1016/j.tics.2006.05.005

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard–easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, *67*(2), 201–221. doi: 10.1006/obhd.1996.0074

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* Garden City, NY: Doubleday.

Todd, P. M., Gigerenzer, G., & ABC Research Group (Eds.). (2012). *Ecological rationality: Intelligence in the world.* New York, NY: Oxford University Press.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*(3), 261–289. doi: 10.1007/s10994-013-5401-4

Van Dolder, D., & van den Assem, M. J. (2018). The wisdom of the inner crowd in three large natural experiments. *Nature Human Behaviour*, *2*(1), 21–26. doi: 10.1038/s41562-017-0247-6

Vickers, D. (1979). *Decision processes in visual perception.* New York, NY: Academic Press.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. doi: 10.1111/j.1467-9280.2008.02136.x

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*(3), 243–268. doi: 10.1002/(SICI)1099-0771(199709)10:3<243::AID-BDM268>3.0.CO;2-M

Wallsten, T. S., & Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, *41*(1), 1–18. doi: 10.1016/S0165-4896(00)00053-6

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(2), 304–309. doi: 10.1073/pnas.1516814112

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*(3), 611–617. doi: 10.1037/0033-2909.110.3.611

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall.

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology: General*, *144*(2), 489–510. doi: 10.1037/xge0000062

3 | When Do Experts Change Their Mind?

Litvinova, A., Herzog, S.M., Hertwig, R., de Zoete, A.,
Ostelo, R., & Kurvers, R.H.J.M.

Abstract

Experts regularly make inconsistent judgments when judging the same case twice. Previous research on experts' inconsistency has largely focused on individual or situational factors. Here we focus directly on the cases themselves. First, using a theoretical model, we study how inconsistency and confidence are affected by how clearly the information in a case points to either the correct or the incorrect decision. Next, we empirically corroborate the model's predictions in two real-world datasets: diagnosticians rating the same mammograms or images of the lower spine twice. Results show that unambiguous cases were associated with highly confident initial decisions that were unlikely to change—independent of whether the experts' consensus decision was correct or incorrect. Furthermore, our results provide simple advice for individuals confronted with two conflicting judgments from a single expert: Choose the more confident decision.

Acknowledgments: We thank Susannah Goss for editing the manuscript.

Introduction

Inconsistency in expert judgments is a prevalent finding in several domains, including medicine (Kirwan, De Saintonge, Joyce, & Currey, 1983; Koran, 1975; Levi, 1989; Ullman & Doherty, 1984), clinical psychology (Little, 1961; Millimet & Greenberg, 1973), neuropsychology (Garb & Schramke, 1996), finance and management (Kahneman, Rosenfield, Gandhi, & Blaser, 2016), agriculture (Trumbo, Adams, Milner, & Schipper, 1962), and weather forecasting (Lusk & Hammond, 1991; Stewart et al., 1989). Such inconsistency is often understood as a source of error (Kahneman et al., 2016) and can have profound consequences; for example, when a physician initially classifies a mass in a breast x-ray as cancerous, but later—when inspecting the same image again—changes her mind and classifies it as benign. Which decision should the patient rely on? Understanding the conditions underlying experts' inconsistency is of key importance—both for developing strategies to improve expert decision making and for giving people advice on what to do when faced with inconsistent expert decisions. Here we address two research questions: When do experts change their mind? And which decision should individuals rely on?

Most studies investigating intraindividual inconsistency focus either on processes within the individual, such as level of experience (Arnhoff, 1954), probabilistic sampling of information (Lewandowsky, Griffiths, & Kalish, 2009; Steyvers, Griffiths, & Dennis, 2006; Vul & Pashler, 2008), and hierarchical hypothesis testing (Brehmer, 1974), or on situational factors, such as time pressure (Rothstein, 1986) and the amount of information available (Einhorn, 1971; Hogarth, 1987). However, one key aspect that has received comparatively little attention is how information within the cases themselves affect inconsistency in experts' judgments (Harvey, 1995; Little, 1961). To the best of our knowledge, none of the previous studies have addressed the interplay between an individual's confidence, consistency, and the ambiguity of a case. This study aims to close this gap in two steps.

First, using a theoretical model (Koriat, 2012a), we study how inconsistency and confidence are affected by how clearly the information in a case points to either the correct or the incorrect decision. We do this by linking an expert's internal consistency (also referred to as "intrarater agreement") to the agreement among a population of experts (also referred to as "interrater agreement"). Next, we empirically test the model's predictions in two real-world datasets: diagnosticians rating the same mammograms (Carney et al., 2012) or images of the lower spine (de Zoete et al., 2002) twice. To preview one major insight not anticipated by previous accounts of expert inconsistency: Cases on which there was clear expert consensus were associated with highly confident initial decisions that were unlikely to change—independent of whether the experts' consensus decision was correct or incorrect.

A Model Linking Inconsistency and Confidence to a Case's Ambiguity

A fundamental process assumed by many models of cognition, judgment, and decision making is that individuals sample evidence from their environment or memory when making a decision (Koriat, 2012a; Lewandowsky

et al., 2009; Pleskac & Busemeyer, 2010; Steyvers et al., 2006; Vul & Pashler, 2008). This sampled evidence determines both the decision and the confidence in that decision (Koriat, 2012a; Kvam & Pleskac, 2016; Pleskac & Busemeyer, 2010). In this view, an individual samples several pieces of evidence ("cues") and selects the option for which there is stronger evidence, and the more clearly the evidence points to the favored option, the more confident an individual will be in the accuracy of that decision. Importantly, in such models, making a second decision about the same case is equivalent to drawing a second sample of evidence. Because the sampling process is probabilistic, the evidence in the second sample can differ from that in the first sample—as can the decision (e.g., "cancer" vs. "no cancer") and the confidence in that decision.

The conditions for inconsistency of repeated judgments and how that inconsistency relates to decision confidence depend on how exactly the sampled evidence determines the decision and the confidence in it. To theoretically investigate this question, we used a simple model embodying the assumptions outlined above to derive key qualitative predictions about the relationship between experts' inconsistency, confidence, and case ambiguity (i.e., how clearly the information contained in the case pointed to one decision or the other). In the Discussion, we show that relaxing the model's assumptions would not change the results of interest and argue that the key predictions would also emerge from more sophisticated models of judgment and decision making, such as evidence accumulation models (e.g., Pleskac & Busemeyer, 2010; Ratcliff & McKoon, 2008).

The model we focus on here is the Self-Consistency Model (SCM; Koriat, 2012a). In the basic version of SCM, a decision maker samples a fixed, odd number of n pieces of evidence ("cues") and chooses the option favored by more cues (i.e., decides between two options using majority voting among cues).[1] Given a probability p of sampling a cue pointing to the correct option (say, "cancer"), the probability P of making a correct decision follows from the binomial distribution:

$$P(p, n) = \sum_{h=m}^{n} \left(\begin{array}{c} n \\ h \end{array} \right) \cdot p^h (1-p)^{n-h}, \tag{1}$$

where $m = \frac{n+1}{2}$ (i.e., the minimum number of cues necessary to decide in favor of the correct decision).

The probability I of making two decisions that are inconsistent is

$$I = P(1-P) + (1-P)P = 2P(1-P), \tag{2}$$

which is maximal ($I = 0.5$) for choices at chance level ($P = 0.5$) and thus by extension for cases that are maximally ambiguous ($p = 0.5$), that is, when every sampled cue is equally likely to point to the correct or the incorrect decision. Conversely, inconsistency is minimal ($I = 0$) for perfectly correct ($P = 1$) and "perfectly" incorrect ($P = 0$) decisions (Figure 1A). This corresponds to cases where all cues point either to the correct decision ($p = 1$; perfectly "kind" cases) or to the incorrect decision ($p = 0$; perfectly "wicked" cases; Hertwig, 2012; Koriat, 2012b). Thus, in SCM, within-expert inconsistency increases the closer a cases's p is to a fair coin flip.[2]

[1] When even numbers of ns are allowed, ties are resolved through a coin flip.

[2] More formally, a case's ambiguity is some monotonically decreasing function of $|p - 0.5|$ (i.e., how close a cases's p is to a fair coin flip).

We will now show how SCM provides a simple, elegant link between the consistency of a single expert's repeated readings of a case and the agreement among experts for that same case. For simplicity, let us assume that all experts sample the same number of cues (i.e., share a common n) and that for any particular case and cue those experts have the same probability p of sampling a cue that points to the correct answer. Although p is not directly observable, according to SCM, the expected proportion of correct decisions $E(P_i(p_i))$ for case i among a population of identical experts is monotonically related to p_i. Empirically, the sample proportion of correct decisions among experts for case i, $\widehat{P}_i$, can be used as a proxy for ordering cases according to their p_i.[3] Thus, assuming SCM, we can use the disagreement among experts (i.e., how close $\widehat{P}_i$ is to 0.5) as an indicator of a case's ambiguity (i.e., how close p_i is to 0.5).

SCM further stipulates that the confidence in a decision increases with the proportion of cues pointing to that decision. In particular, SCM assumes that confidence $\widehat{C}$ is the complement of the sample standard deviation, which depends solely on the proportion of cues pointing to the chosen option ($\widehat{p}$ for correct decisions and $1-\widehat{p}$ for incorrect decisions):

$$\widehat{C} = 1 - \sqrt{\widehat{p}\,(1-\widehat{p})}. \tag{3}$$

Since $\widehat{p} = E(p)$, it follows that confidence is highest for $p = 1$ and $p = 0$ ($\widehat{C} = 1$) and lowest for $p = 0.5$ ($\widehat{C} = 0.5$; Figure 1B)—mirroring the results for an expert's inconsistency (see eq. 2 and Figure 1A).

So when confronted with two inconsistent, conflicting decisions, which should people rely on? The *maximum-confidence slating* (MCS) algorithm (Koriat, 2012b, henceforth "confidence rule") prescribes the decision with the higher confidence. The SCM predicts that, for $p > 0.5$, confidence will be positively correlated with the probability of making a correct decision; for $p < 0.5$, this correlation will be negative.[4]

In sum, SCM predicts:

1. The higher a case's ambiguity (indexed by experts' disagreement among their initial diagnoses), the higher an expert's inconsistency (i.e., the more likely experts will be to make a different diagnosis when judging the same case again; Figure 1A).

2. The higher a case's ambiguity (indexed by experts' disagreement among their initial diagnoses), the less confident an expert will be in her initial diagnosis (Figure 1B).

3. The less confident an expert is in her initial diagnosis, the more likely she will be to change it when judging the same case again (Figure 1C).

4. Considering only cases where an expert makes two inconsistent diagnoses: Relative to sticking with the initial diagnosis, using the confidence rule (i.e., selecting the more confident diagnosis) improves accuracy

[3] Because equation 1 applies to majority voting either over cues or over individuals, we can use Condorcet's Jury Theorem (Condorcet, 1785/1994; Grofman, Owen, & Feld, 1983) to gain insights into how P_i and p_i relate for $n \geq 3$. For example, for $p_i > 0.5 \rightarrow P_i > p_i$; conversely, for $p_i < 0.5 \rightarrow P_i < p_i$. Thus, if we assume that experts sample more than one cue, P_i will be a more extreme version of p_i. Importantly, for any $n \geq 3$, P_i and p_i are identically ordered across a set of cases.

[4] More specifically, equation 1 shows that, for $p > 0.5$, any level of confidence is more likely under the correct than the incorrect decision (and vice versa for $p < 0.5$). To see why, consider that, in equation 1, $\widehat{p} = \frac{h}{n}$. When $p > 0.5$, it follows that $p^h > (1-p)^{n-h}$ and thus the event that a majority of cues (h) point to the correct decision is more likely than the event that the same-sized majority of cues ($n-h$) point to the incorrect decision. For $p < 0.5$, we obtain the opposite result.

for kind items but worsens it for wicked items (i.e., cases where the majority of experts' initial diagnoses were correct vs. incorrect, respectively).

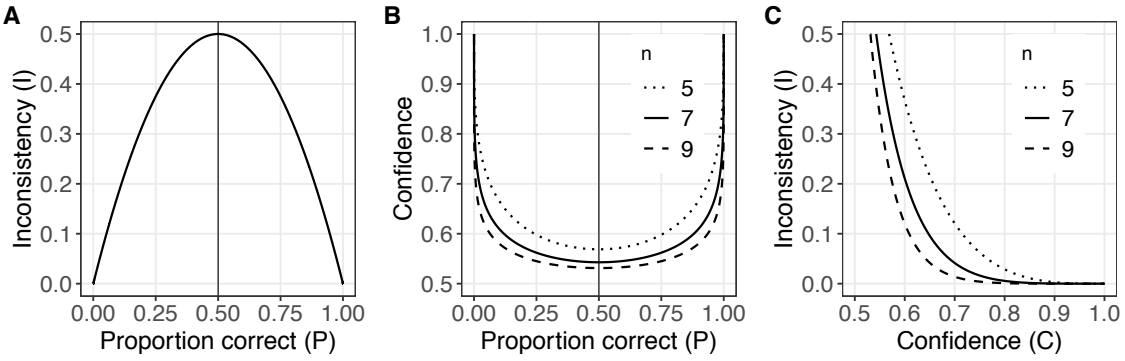


Figure 1. Predictions of the Self-Consistency Model (SCM) on how the proportion of experts who make a correct diagnosis ($\widehat{P}_i$), inconsistency (I; probability of not making the same diagnosis again), and confidence (C) relate to each other for three different values of n (the number of samples retrieved). **(A)** Consistency (I) as a function of $\widehat{P}_i$. Note that the relation between I and $\widehat{P}_i$ does not depend on n (see eq. 2). **(B)** Confidence (C) as a function of $\widehat{P}_i$. **(C)** Inconsistency (I) as a function of confidence (C).

However, these predictions depend on SCM's (Koriat, 2012a) strong assumptions about an expert's judgment process and our additional assumption of complete homogeneity among experts and cues. Specifically, the model assumes that all experts sample the same number of n cues and that, for any particular case and cue, those experts have the same probability p of sampling a cue that points to the correct answer. These assumptions are unlikely to hold for actual expert judgments. Thus, to empirically test the model's predictions and assess how much insight the model provides about the judgments of real experts, we used two real-world high-stakes expert datasets: diagnosticians rating mammograms (Carney et al., 2012) and x-rays of the lower spine (de Zoete et al., 2002). In the following section, we describe the two datasets and how we analyzed them.

Experts' Inconsistency and Confidence in Two Medical Studies

Dataset 1: Radiologists Diagnosing Mammograms (Carney et al., 2012)

Dataset 1 was collected to study the effect of time spent viewing and confidence on diagnostic accuracy in mammography screening. For this study, 572 radiologists were invited to participate, of whom 102 completed all procedures (i.e., phase 1+2, see below). The mammograms used were randomly selected from screening examinations of women aged 40–69 years old. Importantly, the correct diagnosis (cancerous or non-cancerous) for each mammogram was available from follow-up research. In phase 1, each radiologist was randomly assigned to one of four different test sets of 109 mammograms. The radiologists were instructed to interpret the cases as they would in clinical practice. They were informed that the overall cancer rate in their test set was higher than that found in a screened population, but they were not informed of the specific prevalence of positive test results or cancers in their test set. When viewing each case, participants were prompted to identify the most

significant breast abnormality and to decide whether the patient should be recalled for additional workup. The decision to recall constituted a positive test result. Additionally, participants provided a confidence judgment for each assessment ("not at all confident," "not very confident," "neutral," "confident," or "very confident"). Radiologists used either a home or work computer or a laptop provided by the study to complete the task. After an interval of 3–9 months, radiologists were re-invited to rate a second set of 110 mammograms, using an identical procedure as described above. In this phase 2, a proportion of the cases were the same as those presented in phase 1. This information was unknown to the participants. Overall, 58 cases were rated twice by 55 radiologists; of those 58 cases, 46 were rated twice by another 47 radiologists, resulting in 5,352 repeated ratings and either 55 ratings per case (for 12 cases) or 102 ratings per case (for 46 cases) in each assessment phase. All repeated mammograms were noncancer cases (i.e., from women who were cancer-free for at least 2 years after the first mammography). See Carney et al. (2012) for details.

Across all repeated cases, the median of radiologists' accuracies (proportion correct) was median($\widehat{P_j^1}$) = 0.72 for phase 1 and median($\widehat{P_j^1}$) = 0.68 for phase 2.

Dataset 2: Physicians Diagnosing Radiographs of the Lumbosacral Spine (de Zoete et al., 2002)

Dataset 2 was collected to study the diagnostic accuracy of radiologists and chiropractors (total $N = 13$) reading lumbosacral radiographs. Five chiropractors, three chiropractic radiologists, and five medical radiologists participated in the study. Participants' professional experience ranged from 3 to 21 years. 300 radiographs of the lumbosacral spine of adult patients were selected from a general hospital database. For each radiograph, the correct diagnosis was known from follow-up research. The selected radiographs overrepresented "significant abnormalities," such as infections ($n = 7$), malignancies ($n = 15$), fractures ($n = 8$), inflammatory spondylitis ($n = 6$), and spondylolysis ($n = 14$). The set of radiographs was presented in a random order. For each radiograph and each assessment, the physician evaluated whether a significant abnormality was present (in which case immediate referral to a hospital was required) and expressed her confidence in her decision on a two-point scale. Three months later, all 300 radiographs were assessed again by all participants, resulting in 3,900 repeated assessments. See de Zoete et al. (2002) for details.

Across all cases, the median of physicians' accuracy was median($\widehat{P_j^1}$) = 0.86 in phase 1 and median($\widehat{P_j^2}$) = 0.91 in phase 2.

Statistical Analyses

We ran a series of Bayesian mixed-level regression models (using default priors and Bürkner et al. (the R-package *brms*; 2016, version 2.6.0). The models all included group-level intercepts for experts and cases ("random intercepts"). Four chains, each with 4,000 samples (and thinning = 2), were run. The first 2,000 samples were discarded as warm up; thus a total of 4,000 samples were obtained.

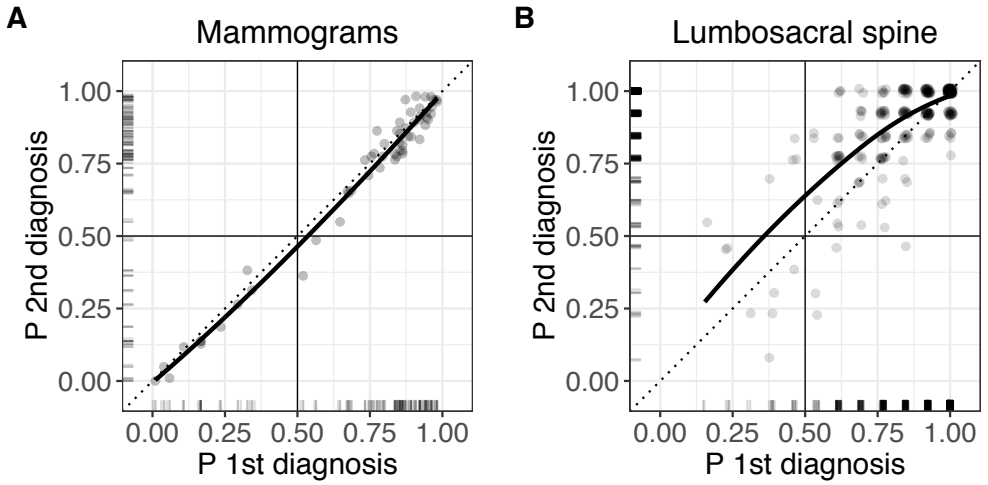


Figure 2. Empirical results on the relations between the proportion of experts who made a correct diagnosis ($\widehat{P}$) **in the first vs. second diagnoses** across cases. Each point represents one case. The rugs on the y- and x-axes show the marginal distributions. The solid curves are LOESS smooths. Panel B employs jittering to avoid overplotting.

Empirical Results

Before presenting the results on our four predictions, we highlight three points. First, although there were 300 radiographs in the spine dataset, each was rated by only 13 experts. Consequently, the estimates for both proportion correct $\widehat{P}_i$ and inconsistency I_i are noisy. In the mammogram dataset, in contrast, up to 102 radiologists rated 58 distinct mammograms, allowing the cases' characteristics to be estimated more reliably. To render our classification of cases in the spine dataset (kind vs. wicked) more reliable, we defined kind cases as $\widehat{P}_i > 0.6$ and wicked cases as $\widehat{P}_i < 0.4$. We thus excluded cases where $0.4 < \widehat{P}_i < 0.6$ in model M7 (Table 2; assessing prediction 4) ; importantly, those cases were retained in all other analyses and all figures (except Figure 4).

Second, experts' average performce was clearly better than chance in both studies, especially in the spine dataset (Figure 2). This implies that there were fewer, and less pronounced, wicked cases than kind cases. As a consequence, predictions 1, 2, and 4 hold with higher certainty for kind than for wicked cases—especially in the spine dataset.

Third, the proportion of experts rendering a correct diagnosis for a mammography case remained largely unchanged across first and second diagnoses (Figure 2A). In contrast, second diagnoses for a spine case had a higher proportion of correct diagnoses (Figure 2B). This latter pattern suggests that—unless second spine diagnoses are also, on average, sufficiently more confident—the confidence rule will likely not outperform the second spine diagnosis.

The results presented in Figure 3A/B corroborate the first prediction: The higher a case's ambiguity (indexed by experts' disagreement among their initial diagnoses), the higher experts' inconsistency—irrespective of whether the experts' consensus opinion for a case was correct or not: In the mammogram dataset (Figure

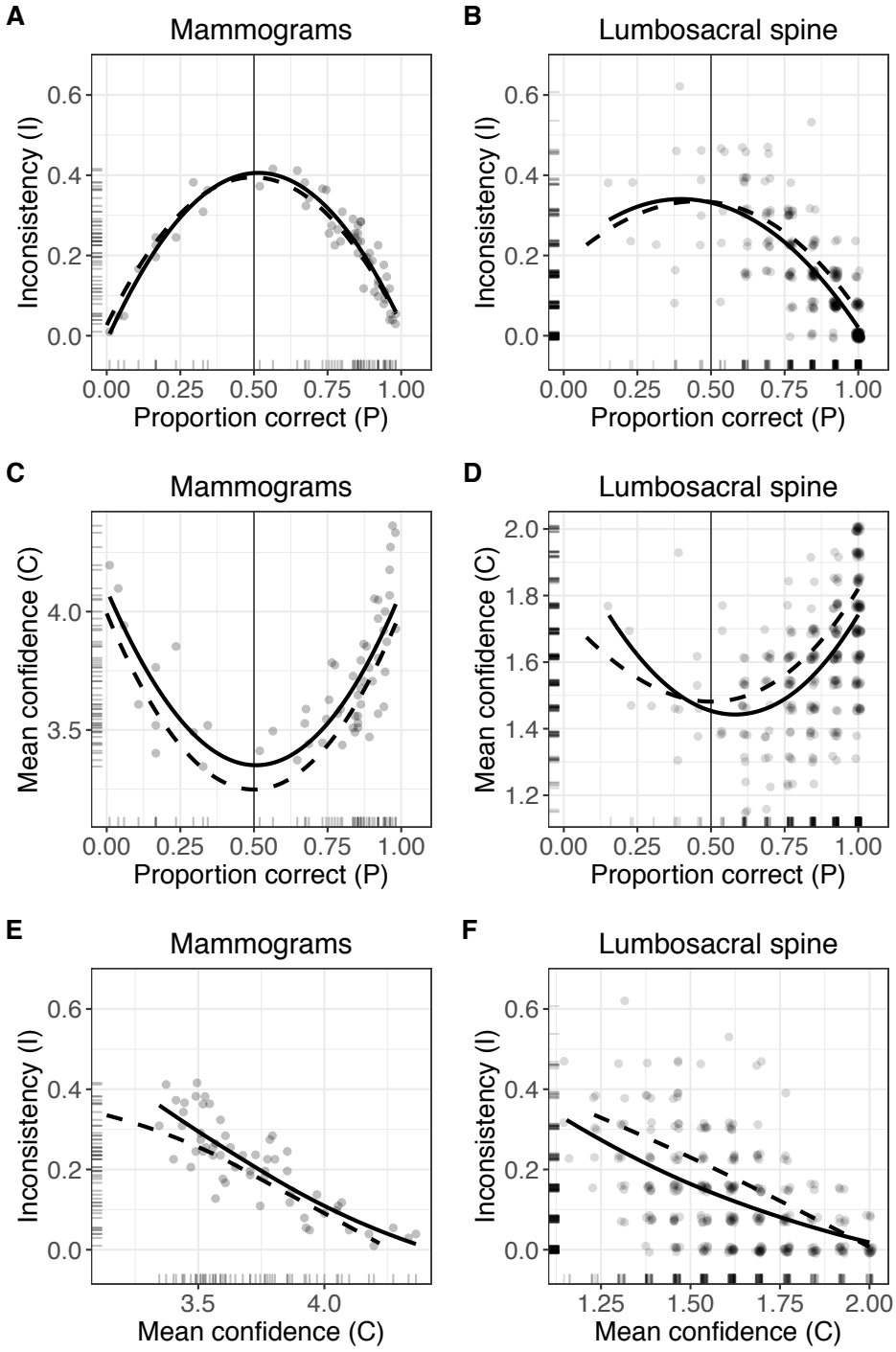


Figure 3. Empirical results on the relations between the proportion of experts who made a correct diagnosis ($\widehat{P}$), inconsistency (I; probability of making not the same diagnosis again), and mean confidence ($\widehat{C}$) in the two datasets. **(A & B)** Inconsistency (I) as a function of $\widehat{P}$. **(C & D)** Mean confidence ($\widehat{C}$) as a function of $\widehat{P}$. **(E & F)** Inconsistency (I) as a function of mean confidence ($\widehat{C}$). Each point represents one case and its coordinates represent $\widehat{P}$ and $\widehat{C}$ from initial diagnoses. The rugs on the y- and x-axes show the marginal distributions. The solid curves are LOESS smooths. For comparison, the dashed curves show the smooths resulting from using $\widehat{P}$ and $\widehat{C}$ from the second diagnoses (the corresponding points are not shown). Panels B, D, and F employ jittering to avoid overplotting.

	Mammography			Lumbosacral spine		
Parameter	Estimate	95% CI		Estimate	95% CI	
M1: Inconsistency (intercept-only model)						
Intercept	−1.51	−1.76	−1.27	−2.31	−2.71	−1.90
sd(expert)	0.47	0.37	0.59	0.62	0.40	1.06
sd(case)	0.77	0.61	0.98	0.95	0.78	1.13
M2: Inconsistency vs. case ambiguity (Prediction 1): $I \sim (\widehat{P} - 0.5) + (\widehat{P} - 0.5)^2$						
Intercept	−1.50	−1.63	−1.37	−2.39	−2.79	−2.01
$(\widehat{P} - 0.5)$	−1.61	−8.48	5.53	−56.97	−63.38	−50.69
$(\widehat{P} - 0.5)^2$	−49.09	−56.49	−42.63	−27.41	−33.14	−21.60
sd(expert)	0.47	0.37	0.59	0.64	0.40	1.08
sd(case)	0.17	0.03	0.30	0.09	0.01	0.28
M3: Confidence (intercept-only model)						
Intercept	3.73	3.61	3.85	1.62	1.45	1.78
sd(expert)	0.46	0.40	0.54	0.29	0.20	0.46
sd(case)	0.25	0.21	0.31	0.16	0.15	0.18
M4: Confidence vs. case ambiguity (Prediction 2): $C \sim (\widehat{P} - 0.5) + (\widehat{P} - 0.5)^2$						
Intercept	3.73	3.63	3.82	1.62	1.45	1.80
$(\widehat{P} - 0.5)$	2.97	−0.30	6.33	5.45	4.33	6.57
$(\widehat{P} - 0.5)^2$	13.95	10.57	17.11	4.16	3.04	5.36
sd(expert)	0.46	0.40	0.54	0.29	0.20	0.48
sd(case)	0.16	0.13	0.20	0.12	0.11	0.14
M5: Inconsistency vs. confidence (Prediction 3): $I \sim (C - 1)$						
Intercept	0.16	−0.20	0.51	−1.39	−1.72	−1.07
$(C - 1)$	−0.63	−0.74	−0.52	−1.66	−1.92	−1.41
sd(expert)	0.54	0.43	0.67	0.44	0.27	0.76
sd(case)	0.62	0.49	0.81	0.78	0.60	0.96

Table 1. Bayesian mixed-level regression models for predictions 1, 2, and 3 in the mammography and lumbosacral spine datasets. The three models of inconsistency (M1, M2, and M4) are logistic regression models and thus the parameters indicate (changes in) log odds. The two models of confidence (M3 and M5) are linear models (i.e., identity link). Posterior distributions of parameters are summarized by their posterior median (*Estimate*) and 95% credible interval. *sd(expert)* and *sd(case)* show the standard deviations of the group-level distribution of the intercept for experts and cases, respectively. $(\widehat{P} - 0.5)$ and $(\widehat{P} - 0.5)^2$ in models M2 and M4 are the linear and quadratic polynomial contrasts of the 0.5-centered proportion of correct diagnoses per case; this means that the intercept in those models predicts the value of the dependent variable for a maximally ambiguous case ($\widehat{P} = 0.5$; because for $\widehat{P} = 0.5$, $(\widehat{P} - 0.5) = (\widehat{P} - 0.5)^2 = 0$). $(C - 1)$ in model M5 is the linear effect of confidence, re-coded so that the intercept indicates the inconsistency at the lowest confidence level in both datasets (i.e., $C = 1$; this is because for $C = 1$, $(C - 1) = 0$ corresponds to the lowest possible confidence rating).

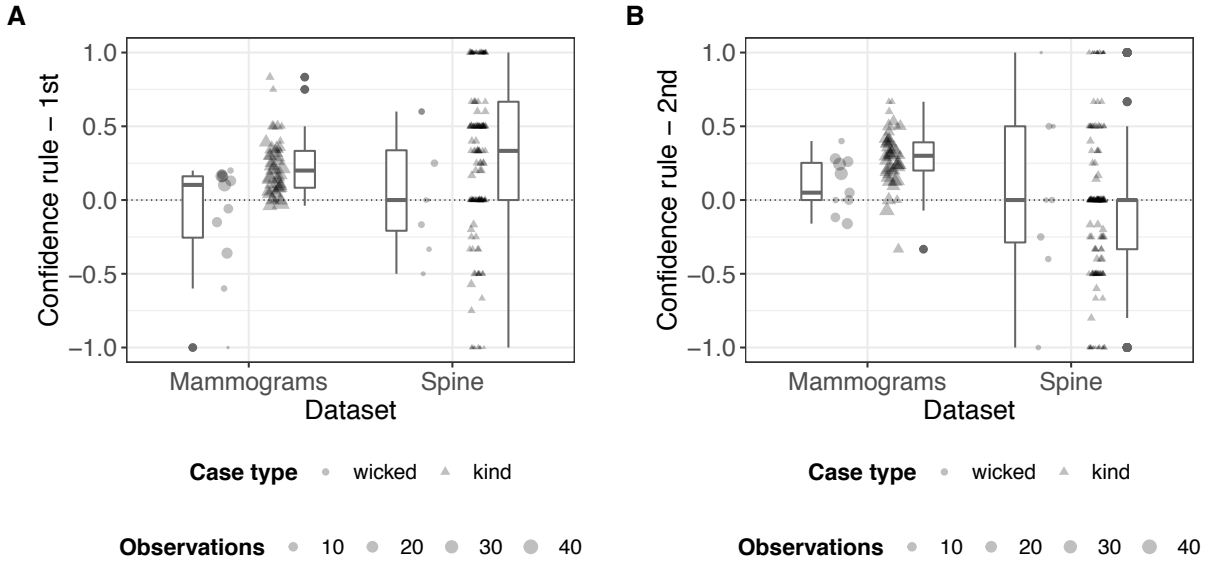


Figure 4. Empirical results comparing the accuracy of the confidence rule to the accuracy of first (A) and second (B) diagnoses for cases where experts were inconsistent, separately for kind and wicked cases and both datasets (mammography and lumbosacral spine). Positive values on the y-axes indicate that the confidence rule outperformed first diagnoses (A) or second diagnoses (B), respectively. Cases are shown as jittered shapes and are summarized by boxplots. The size of the shapes indicates the number of experts contributing to each case.

3A), experts became more consistent the more strongly they initially favored either the correct or the incorrect diagnosis. In the spine dataset, the predicted pattern is also clearly visible for kind cases (Figure 3B). However, the results for the few wicked items are less clear (Figure 3B). The regression models M2 (Table 1) show a clearly negative quadratic term in both datasets, corroborating the visual impression from Figure 3A/B. Comparing the standard deviations of the group-level intercepts for experts and cases in model M1 (intercept-only model) shows that cases differed more strongly in inconsistency than experts. With respect to case ambiguity (model M2), the standard deviation for cases is reduced by a factor of 5 in the mammography dataset and by a factor of 11 in the spine dataset, highlighting how much variance in inconsistency can be explained by a case's ambiguity.

The results presented in Figure 3C/D corroborate the second prediction: The higher a case's ambiguity (indexed by experts' disagreement among their initial diagnoses), the less confident experts are in their initial diagnoses—again, irrespective of whether the experts' consensus opinion for a case was correct or not. The regression models M4 (Table 1) show a clearly positive quadratic term in both datasets, corroborating the visual impression from Figure 3C/D.

The results shown in Figure 3E/F corroborate the third prediction: The less confident experts are in their initial diagnosis, the more likely they will be to change it when judging the same case again. The regression models M5 (Table 1) show a clearly negative linear term in both datasets, corroborating the visual impression from Figure 3E/F.

Our final results relate to the fourth prediction: When considering only cases where an expert made two

	Mammography			Lumbosacral spine		
Parameter	Estimate	95% CI		Estimate	95% CI	
M6: Confidence rule vs. first/second diagnoses						
Intercept	0.78	0.66	0.91	0.38	0.19	0.58
Diagnosis 1	−0.58	−0.76	−0.41	−1.04	−1.30	−0.78
Diagnosis 2	−0.98	−1.15	−0.81	0.29	0.03	0.55
sd(expert)	0.03	0.00	0.11	0.10	0.01	0.30
sd(case)	0.04	0.00	0.12	0.05	0.00	0.15
M7: Confidence rule and kind vs. wicked cases (Prediction 4)						
Intercept	0.97	0.82	1.12	0.51	0.28	0.77
Wicked	−0.61	−0.92	−0.28	−0.76	−1.87	0.28
Diagnosis 1	−0.82	−1.02	−0.61	−1.11	−1.42	−0.79
Diagnosis 1 × Wicked	0.66	0.23	1.09	−0.20	−2.02	1.48
Diagnosis 2	−1.13	−1.33	−0.93	0.10	−0.21	0.41
Diagnosis 2 × Wicked	0.56	0.12	0.98	1.77	0.17	3.53
sd(expert)	0.03	0.00	0.12	0.13	0.01	0.37
sd(case)	0.03	0.00	0.11	0.05	0.00	0.18

Table 2. Bayesian mixed-level regression models for prediction 4 in the mammography and lumbosacral spine datasets, considering only cases where an expert's two diagnoses for the same case differed. Both models M6 and M7 are logistic regression models and thus the parameters indicate (changes in) log odds. The decision of the confidence rule is the reference level, that is, *Diagnosis 1* and *Diagnosis 2* in model M6 indicate the change in accuracy (in log odds) from the confidence rule (*Intercept*) to the first or second diagnosis, respectively. In model M7, kind cases constitute the reference level, that is, *Wicked* indicates for kind cases the change in accuracy (in log odds) from the confidence rule (*Intercept*) to wicked cases. Then *Diagnosis 1* and *Diagnosis 2* indicate for kind cases the change in accuracy (in log odds) from the confidence rule (*Intercept*) to the first or second diagnosis, respectively. The interaction terms (*Diagnosis 1* × *Wicked* and *Diagnosis 2* × *Wicked*) show whether the type of case (kind vs. wicked) moderates the differences between the confidence rule and first and second diagnoses, respectively. Posterior distributions of parameters are summarized by their posterior median (*Estimate*) and 95% credible interval. *sd(expert)* and *sd(case)* show the standard deviations of the group-level distribution of the intercept for experts and cases, respectively.

inconsistent diagnoses, we found that, relative to sticking with the initial diagnosis, using the confidence rule (i.e., selecting the more confident diagnosis) improves accuracy for kind items but worsens it for wicked items (i.e., cases where the majority of experts' initial diagnoses were correct vs. incorrect, respectively). Figure 4 shows that for kind cases the confidence rule was more accurate than either the first or second diagnosis in the mammography dataset, but only more accurate than the first diagnosis in the spine dataset. The results for wicked cases were less consistent with our fourth prediction. In the mammography dataset, the difference in performance between the confidence rule and first and second diagnoses was reduced; however, the confidence rule did not, as predicted, perform worse than first and second diagnoses. In the spine dataset, a similar result was found when we compared the confidence rule against the first diagnosis, but comparing the confidence rule against the second diagnosis did not reveal a clear difference. Model M7 (Table 2) corroborates these observations.

General Discussion

When do experts change their mind? Previous research on inconsistency has focused largely on individual factors (e.g., Lewandowsky et al., 2009; Steyvers et al., 2006; Vul & Pashler, 2008) or situational factors

(Einhorn, 1971; Hogarth, 1987; Rothstein, 1986). Here we focus directly on the cases themselves. First, using the SCM (Koriat, 2012a), we studied how inconsistency and confidence are affected by how clearly the information in a case points to either the correct or the incorrect decision (a case's ambiguity, indexed by experts' disagreement among their initial diagnoses). Next, we empirically confirmed three of the model's four key predictions in two real-world datasets: diagnosticians rating the same mammograms (Carney et al., 2012) or images of the lower spine (de Zoete et al., 2002) twice. We found that the higher a case's ambiguity, the higher experts' inconsistency (prediction 1) and the lower their confidence in their initial diagnosis (prediction 2)—irrespective of whether the experts' consensus opinion for a case was correct or not. The first two predictions imply that the more confident an expert is in her initial diagnosis, the less likely she will be to change her diagnosis when judging the same case again (prediction 3), irrespective of whether the experts' consensus opinion was correct or not. Together, these first three results imply that a highly confident or unchanged diagnosis is first and foremost an indicator for how strong the consensus among experts is, and an indicator of accuracy only to the extent that most cases in the domain of interest are kind.

Finally, when an expert's two diagnoses were inconsistent, using the confidence rule (i.e., selecting the more confident diagnosis) improved accuracy (relative to sticking with the initial diagnosis). However, this fourth prediction was empirically corroborated only for kind cases and only partially corroborated for wicked cases—although other results for wicked cases were consistent with predictions 1 and 2 (see the left sides of panels A–D in Figure 3). These latter, mixed findings might be explained by systematic differences between first and second diagnoses in terms of accuracy and confidence judgments, especially for the spine dataset. For example, second spine diagnoses were more accurate than first ones (Figure 2B). Future research should explore the implications of such systematic differences. Importantly, however, model M6 (Table 2) shows that across all cases the confidence rule outperformed both first and second diagnoses in the mammography dataset and first, but not second, diagnoses in the spine dataset. Because decision makers cannot, in practice, tell in advance whether a particular case is kind or wicked (Koriat, 2017), this result means that the confidence rule has clear practical merit from an applied perspective. It thus implies the following prescription: Unless you suspect that experts perform worse than chance in the domain of interest, rely on the more confident of two conflicting judgments from an individual expert.

Linking Inconsistency to a Case's Ambiguity

In the past, research has primarily studied inconsistency from an internal individual perspective, which explains inconsistency as a consequence of unreliable processing of information. Studies reported, for example, that an individual's judgments become less reliable as the amount of available information increases (Einhorn, 1971; Hogarth, 1987) because their capacity to process that information decreases (Faust, 1986; Sen & Boe, 1991). Other studies showed that unreliability in judgment can often be attributed to a lack of cognitive control—how acquired knowledge is used—rather than a lack of knowledge (Hammond & Summers, 1972). Harvey (1995) reviewed further reasons for inconsistency in decisions, such as overload in working memory, learning

correlations instead of functions, or reproduction of noise. In contrast to these individual factors, external-task related factors contributing to experts' consistency, such as the predictability of the environment (i.e., the degree to which cues allow the outcome to be predicted) have been rarely studied. A set of studies has shown that, as a task becomes less predictable, individuals make less consistent judgments (Brehmer, 1976; Camerer, 1981; Harvey, 1995). Our results are largely consistent with these previous findings. Importantly, however, none of these previous studies made the connection between an individual's confidence, consistency, and the ambiguity of a case. Below we discuss two contributions that our perspective on expert inconsistency offers.

First, in our reading of the literature, previous accounts of experts' inconsistency explicitly or implicitly assume or predict that consistency increases as the accuracy of a judgment increases. In stark contrast to this assumption, our results show that this pattern is mirrored at chance level: For cases that experts tend to judge incorrectly, consistency starts to increase again the more experts agree on the incorrect diagnosis. Furthermore, our results showed that confidence tracks consistency, but because confidence tracks the ambiguity of a case and not accuracy per se (Koriat, 2012a), confidence's ability to predict accuracy and consistency strongly depends on the environment (i.e., the distribution of ambiguity across the cases). If there are only kind cases (i.e., cues tend to point to the correction decision), confidence strongly predicts that a diagnosis is accurate and will not change. The more wicked cases there are (i.e., cues tend to point to the wrong decision), the more these relations dilute. In the extreme case of a domain dominated by wicked cases (in which experts, on average, tend to make wrong decisions), the two relations dissociate: Experts' confidence is then *negatively* related to accuracy, but still positively related to consistency—and being consistent in a wicked environment means confidently sticking to the wrong diagnosis.

Second, previous accounts have focused on differences in consistency among experts or in different task conditions (e.g., time pressure). Our perspective predicts that the cases themselves can differ markedly in how consistently they are diagnosed by any expert. Importantly, as our results have shown, these consistency differences among cases can be even larger than those observed among the experts themselves and can be explained to a large degree by a case's level of ambiguity.

Would Relaxing our SCM's Assumptions Lead to Different Predictions?

In this study, we used the SCM (Koriat, 2012a) to gain insights into when experts are inconsistent and what to do as a decision maker when faced with inconsistent advice from the same expert. Our implementation of the SCM assumes that all experts sample the same number of n cues and that for any particular case and cue those experts have the same probability p of sampling a cue that points to the correct answer. We argue that relaxing those assumptions will, in general, not change the four key predictions; it will affect how exactly the probability of a correct decision, P, depends on p and n, but the qualitative implications of the distinction between kind cases ($p > 0.5$) and wicked cases ($p < 0.5$) will remain unchanged. In addressing this point, we can benefit from the fact that SCM's decision process amounts to majority voting among cues; we

can therefore apply insights from more general research on majority voting. For example, similar conclusions follow if, within an expert, the cues' probabilities p_i are not identical but symmetrically distributed around p_i (Grofman et al., 1983), or if the retrieval of cues is allowed to be dependent (e.g., retrieving cues pointing to one option increases the likelihood that further cues point to that same option; Grofman et al., 1983; Ladha, 1992, 1995). As another example, under very general conditions, as the number of cues retrieved, n, increases, the probability of a correct decision, P, will increase for kind cases ($p > 0.5$) and decrease for wicked cases ($p < 0.5$; Grofman et al., 1983; Ladha, 1992, 1995). As a consequence, keeping everything else constant, consistency should increase as more cues are retrieved (see eq. 2); the variation in cases' ambiguity will be most pronounced for small ns, whereas for large ns all cases will be clearly diagnosed either correctly or incorrectly. Furthermore, assuming that experts sample different numbers of cues implies that, for the same case, experts with larger ns will be more consistent than experts with smaller ns.

Would Different Models Make Different Predictions?

In this study, we focused on the SCM (Koriat, 2012a) as a simple model linking accuracy, confidence, and consistency, but we argue that a broad family of models make qualitatively similar predictions. However, to the best of our knowledge, the ability of these models to gain insights into expert consistency and the role of wicked cases has not yet been explored. For example, in the diffusion decision model (Ratcliff & McKoon, 2008), a prominent example of the family of evidence accumulation models, case ambiguity is reflected in the drift rate, which represents the average speed with which an individual accumulates evidence that stochastically drifts to one of two decision boundaries (e.g., correct vs. incorrect answer). Keeping everything else constant, a reduction in the drift toward zero implies increasingly more ambiguous cases, which are predicted to be associated with lower accuracy, longer response times, and lower confidence (Pleskac & Busemeyer, 2010; Ratcliff & McKoon, 2008). Drift rates below zero represent wicked cases, where the evidence tends to accumulate to the wrong decision boundary. Importantly, an increasingly negative drift rate corresponds to increasingly less ambiguous and more wicked cases, which are predicted to be associated with lower accuracy, but *shorter* response times, and *higher* confidence—thus qualitatively mirroring the predictions from the SCM. More generally, any model that assumes or implies the following two relations should predict qualitatively similar results: A particular decision becomes more likely and is rendered more confidently the more clearly the relevant information points to that decision. If this claim sounds both grandiose and self-evident, it is because those two relations are fundamental to many, if not most, psychological and normative models of decision making. The crux of the matter is, of course, what exactly it means in a particular model that a case is kind or wicked and clear cut or ambiguous. This will depend both on the statistical structure of the cue–criterion and intercue relationships, and on the decision strategy applied. Future research should map out which decision environments result from combinations of which decision strategies and cue structures.

Conclusion

Inconsistency in expert judgment is a common finding in various domains, including medicine, finance and management, and weather forecasting. Most previous studies have investigated inconsistency from an individual or situational perspective, leading to methods to improve information processing within individuals. In this study, we connected—theoretically and empirically—the ambiguity of the case with the confidence and inconsistency of the expert. For individuals confronted with inconsistent judgments from a single expert, we advise the following: Unless there is reason to believe that the expert performs below chance, rely on the more confident judgment.

References

Arnhoff, F. N. (1954). Some factors influencing the unreliability of clinical judgments. *Journal of Clinical Psychology*, *10*(3), 272–275. doi: 10.1002/1097-4679(195407)10:3<272::AID-JCLP2270100319>3.0.CO;2-V

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27. doi: 10.1016/0030-5073(74)90002-6

Brehmer, B. (1976). Note on clinical judgment and the formal characteristics of clinical tasks. *Psychological Bulletin*, *83*(5), 778–782. doi: 10.1037/0033-2909.83.5.778

Bürkner, P.-C., et al. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance*, *27*(3), 411–422. doi: 10.1016/0030-5073(81)90031-3

Carney, P. A., Bogart, T. A., Geller, B. M., Haneuse, S., Kerlikowske, K., Buist, D. S., ... Miglioretti, D. L. (2012). Association between time spent interpreting, level of confidence, and accuracy of screening mammography. *American Journal of Roentgenology*, *198*, 970–978.

Condorcet, N. C. (1785/1994). Essay on the application of probability analyses to decisions returned by a plurality of people. In *Condorcet: Foundations of social choice and political theory* (pp. 11–36). Brookfield,VT: Edward Elgar. (Original work published 1785)

de Zoete, A., Assendelft, W. J., Algra, P. R., Oberman, W. R., Vanderschueren, G. M., & Bezemer, P. D. (2002). Reliability and validity of lumbosacral spine radiograph reading by chiropractors, chiropractic radiologists, and medical radiologists. *Spine*, *27*(17), 1926–1933. doi: 10.1097/01.BRS.0000025722.90766.35

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance*, *6*(1), 1–27. doi: 10.1016/0030-5073(71)90002-X

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology: Research and Practice*, *17*(5), 420–430. doi: 10.1037/0735-7028.17.5.420

Garb, H. N., & Schramke, C. J. (1996). Judgment research and neuropsychological assessment: A narrative review and meta-analyses. *Psychological Bulletin*, *120*(1), 140–153. doi: 10.1037/0033-2909.120.1.140

Grofman, B., Owen, G., & Feld, S. (1983). Thirteen theorems in search of the truth. *Theory and Decision*, *15*(3), 261–278. doi: 10.1007/bf00125672

Hammond, K. R., & Summers, D. A. (1972). Cognitive control. *Psychological Review*, *79*(1), 58–67.

Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes*, *63*(3), 247–263. doi: 10.1006/obhd.1995.1077

Hertwig, R. (2012). Tapping into the wisdom of the crowd—with confidence. *Science*, *336*(6079), 303–304.

Hogarth, R. M. R. M. (1987). *Judgment and choice: The psychology of decision, 2nd edition*. Chichester, WS: Wiley and Sons.

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*, *94*(10), 38–46.

Kirwan, J., De Saintonge, D. C., Joyce, C., & Currey, H. (1983). Clinical judgment in rheumatoid arthritis. i. rheumatologists' opinions and the development of 'paper patients'. *Annals of the Rheumatic Diseases*, *42*(6), 644–647.

Koran, L. M. (1975). The reliability of clinical methods, data and judgments. *New England Journal of Medicine*, *293*(14), 695–701. doi: 10.1056/NEJM197510022931405

Koriat, A. (2012a). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. doi: 10.1037/a0025648

Koriat, A. (2012b). When are two heads better than one and why? *Science*, *336*(6079), 360–362. doi: 10.1126/science.1216549

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. doi: 10.1002/bdm.2024

Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition*, *152*, 170–180. doi: 10.1016/j.cognition.2016.04.008

Ladha, K. K. (1992). The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 617–634.

Ladha, K. K. (1995). Information pooling through majority-rule voting: Condorcet's jury theorem with correlated votes. *Journal of Economic Behavior & Organization*, *26*, 353–372.

Levi, K. (1989). Expert systems should be more accurate than human experts: evaluation procedures from human judgement and decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*(3), 647–657. doi: 10.1109/21.31070

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*(6), 969–998. doi: 10.1111/j.1551-6709.2009.01045.x

Little, K. B. (1961). Confidence and reliability. *Educational and Psychological Measurement*, *21*(1), 95–101.

Lusk, C. M., & Hammond, K. R. (1991). Judgment in a dynamic task: Microburst forecasting. *Journal of Behavioral Decision Making*, *4*(1), 55–73. doi: 10.1002/bdm.3960040105

Millimet, C. R., & Greenberg, R. P. (1973). Use of an analysis of variance technique for investigating the differential diagnosis of organic versus functional involvement of symptoms. *Journal of Consulting and Clinical Psychology*, *40*(2), 188–195. doi: 10.1037/h0034568

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. doi: 10.1037/a0019737

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. doi: 10.1162/neco.2008.12-06-420

Rothstein, H. G. (1986). The effects of time pressure on judgment in multiple cue probability learning. *Organizational Behavior and Human Decision Processes*, *37*(1), 83–92. doi: 10.1016/0749-5978(86)90045-2

Sen, T., & Boe, W. J. (1991). Confidence and accuracy in judgements using computer displayed information. *Behaviour & Information Technology*, *10*(1), 53–64. doi: 10.1080/01449299108924271

Stewart, T. R., Moninger, W. R., Brady, R. H., Merrem, F. H., Stewart, T. R., & Grassia, J. (1989). Analysis of expert judgment in a hail forecasting experiment. *Weather and Forecasting*, *4*(1), 24–34. doi: 10.1175/1520-0434(1989)004<0024:AOEJIA>2.0.CO;2

Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, *10*(7), 327–334. doi: 10.1016/j.tics.2006.05.005

Trumbo, D., Adams, C., Milner, M., & Schipper, L. (1962). Reliability and accuracy in the inspection of hard red winter wheat. *Cereal Science Today*, *7*, 62–71.

Ullman, D. G., & Doherty, M. E. (1984). Two determinants of the diagnosis of hyperactivity: The child and the clinician. *Advances in Developmental & Behavioral Pediatrics*, *5*, 167–219.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. doi: 10.1111/j.1467-9280.2008.02136.x

4 | Cognitive Dependencies in Sequential Diagnostic Reasoning Tasks

Litvinova, A., Herzog, S. M., Kurvers, R. H. J. M., & Hertwig, R.

Abstract

In sequential diagnostic reasoning tasks, the order of the evidence a person encounters can influence their final diagnosis. We hypothesized that sequential diagnostic procedures such as categorizing a sequence of cues in a skin lesion can induce cognitive dependencies in the classification of cues. Using a signal detection model we investigated three cognitive dependency processes: (i) the *confirmatory response effect*, which biases an individual towards the hypothesis corresponding to their first cue assessment, (ii) the *confirmatory evidence accumulation* process, which influences an individual's evidence accumulation towards the hypothesis corresponding to their first cue assessment, and (iii) the *relevance effort effect*, which influences an individual's effort depending on the relevance of the consecutive cues for the final diagnosis. To reduce the dependency processes, we designed a condition in which the assessment of each cue is temporally isolated and compared it to the common sequential diagnostic procedure, a grouped process where all cues in one case are fully assessed before turning to the next case. We tested both conditions in two environments with either uncorrelated or moderately positively correlated cues. All hypotheses were preregistered. Results suggest the presence of the confirmatory response effect and the confirmatory evidence accumulation process in the correlated environment. Isolating cue assessments did not reduce the strength of the dependency processes. Future research should investigate the downstream consequences of dependent cue assessments on the accuracy of the final diagnosis, identify beneficial levels of dependency and study the extent to which individuals adapt to the statistical environment.

Acknowledgments: We thank Deborah Ain for editing the manuscript, Caroline Graf for coding the experiment, and Jann Wäscher for assistance with data collection. Furthermore, we thank Tim, J. Pleskac and Alexander Fengler for their advice on the initial DDM model.

Introduction

In sequential diagnostic reasoning tasks the order of the evidence a person encounters can influence their final diagnosis. In medical scenarios with multiple possible hypotheses around a diagnosis, ambiguous pieces of evidence are distorted toward the hypothesis that corresponds with the piece of evidence that was encountered first (Kostopoulou, Mousoulis, & Delaney, 2009; Lange, Thomas, & Davelaar, 2012; Rebitschek, Bocklisch, Scholz, Krems, & Jahn, 2015), suggesting that once an individual has an emerging judgment, they tend to distort additional information such that it coheres with that judgment. Relatedly, individuals are inclined to bias their final diagnosis towards the hypothesis they initially formed (Rebitschek, Krems, & Jahn, 2015). Furthermore, when individuals make judgments under incomplete evidence they try to infer the presence of evidence using irrelevant information and then use that inferred evidence to explain their judgment (Johnson, Rajeev-Kumar, & Keil, 2016), suggesting that the human mind will fill in gaps in a diagnostic reasoning process. In the health care sector, medical error is estimated to be the third leading cause of death in the United States (Makary & Daniel, 2016). Given the growing literature revealing not only systemic but also cognitive mechanisms behind diagnostic errors, this paper investigates the underlying cognitive processes in sequential diagnostic decision-making and explores how to improve the decision-making process.

The "three-point" checklist of dermoscopy is a sequential diagnostic tool used by dermatologists to detect skin cancer. The checklist prescribes counting the presence of three cues (blue/white color, atypical network, and asymmetry) and issues a malignant diagnosis if two or more cues are present; if one or no cues are present, the lesion is considered benign (Argenziano et al., 2003; Zalaudek et al., 2006). This checklist belongs to a general class of heuristic decision-making strategies known as *tallying*. Tallying is a decision rule that counts the number of cues supporting one alternative over the other and compares this tally to a threshold in order to make a decision or categorization (Dawes, 1979; Martignon, Katsikopoulos, & Woike, 2008). Experimental studies and simulations using natural and synthetic domains have shown that tallying can perform extremely well compared to more costly strategies that also consider the weights of cues, such as multiple regression models (Czerlinski, Gigerenzer, & Goldstein, 1999; Kattah, Talkad, Wang, Hsieh, & Newman-Toker, 2009; Martignon et al., 2008; McCammon & Hägeli, 2007). In many domains, however, once an accurate decision rule has been developed, individuals still need to assess the cue values (e.g., "Does this skin lesion have an atypical network or not?") before they can apply a decision rule such as tallying (e.g., "If two or more cues are present, diagnose as malignant; otherwise, diagnose as benign"). Extant research in psychology, cognitive science, and judgment and decision making, in contrast, has focused on how individuals learn decision strategies when the cues themselves are simply presented as a given and thus there is no need to assess them (e.g., a geometric shape has one of two colors, or a city has an airport or not; Ashby & Maddox, 2005; Gigerenzer, Hertwig, & Pachur, 2011; Kruschke, 2008; Payne, Bettman, & Johnson, 1993). Similarly, in the above reviewed studies on biases and distortion of information in sequential diagnostic reasoning (Johnson et al., 2016; Kostopoulou et al., 2009; Rebitschek, Bocklisch, et al., 2015; Rebitschek, Krems, & Jahn, 2015) cues were presented; they did not need to be assessed. In order to investigate individuals' cognitive processes when applying a categorization

strategy and explore ways to improve the overall decision-making process, we examined whether tallying induces cognitive dependency processes that influence the classification of sequentially assessed cues and tested whether dependency can be reduced by redesigning the diagnostic procedure—in this case, by breaking up the sequence in order to temporally separate the assessment of cues.

Cognitive Dependency Processes

Let us return to the three-point checklist, which issues a malignant diagnosis if two or more cues are present and a benign diagnosis otherwise. Suppose a dermatologist evaluating a skin lesion assesses the first cue as present. Could that first assessment influence the dermatologist's consecutive cue assessments? For example, does the awareness that only one more present cue is needed for a malignant diagnosis affect the cognitive processes in assessing further cues?

To explore whether the sequential diagnostic procedure of a tallying rule, such as the three-point checklist, can induce dependency in the assessments of cues, we focused on three mutually compatible cognitive dependency processes. First, the *confirmatory response effect* states that the first cue assessment (e.g., the dermoscopic cue A is present) could bias an individual's response threshold in assessing the second cue towards the hypothesis that corresponds with the first cue assessment (Germar, Albrecht, Voss, & Mojzisch, 2016; Germar, Schlemmer, Krug, Voss, & Mojzisch, 2014; Rebitschek, Krems, & Jahn, 2015). For example, if the first cue is judged to be present, this assessment contributes positively to the "malignacy" hypothesis because it increases the tally by one; therefore a positive assessment of the first cue might bias the response threshold such that the dermatologist is more likely to judge the second cue as present.

Second, the *confirmatory evidence accumulation process* states that the first cue assessment could bias evidence accumulation towards the hypothesis that corresponds with the first cue assessment. For example, if the first cue was assessed as present, this assessment might direct our dermatologist's attention to search for patterns that correspond with the hypothesis of the first cue assessment. If the assessed value of the first cue is congruent with the true state of the second cue, confirmatory evidence accumulation would increase the efficiency of extracting evidence from the stimulus and decrease the efficiency otherwise (Germar et al., 2016, 2014; Talluri, Urai, Tsetsos, Usher, & Donner, 2018).

Third, the *relevance effort effect* states that the effort spent on the second cue assessment depends on whether that assessment could change the overall decision. Imagine that there are only two cues and the tallying threshold for issuing a malignant diagnosis is two (i.e., both cues are present). If the first cue is judged as absent, then the second cue assessment cannot change the final diagnosis—even if the second cue is judged to be present, the final diagnosis would be "benign" because only one cue is present and the threshold has not been met. Thus, the effort in judging the second cue may decrease, thereby diminishing the decision maker's ability to discriminate whether a cue is absent or present. In contrast, if the first cue is judged as present, the second cue judgment is critical for the overall decision (if judged to be present, the diagnosis would be "malignant"; if judged to be absent, the diagnosis would be "benign") and effort should stay high and contribute

positively to the decision maker's ability to discriminate.

The three processes will not only affect the accuracy of second cue assessments, but also have downstream consequences for the accuracy of the final diagnosis. For example, if a dermatologist wrongly assesses the first cue as present, then a confirmatory response effect would make them more likely to evaluate the second cue assessment as present. If the first two cues are assessed as present (i.e., a tally of 2) but the true tally is actually 0, they will make a wrong final diagnosis, classifying the lesion as malignant although it is benign. Research on the wisdom of the (inner) crowd has shown that gains from aggregating judgments are larger, the more independent the judgments, and hence their errors are (Herzog & Hertwig, 2009; Larrick, Mannes, & Soll, 2012; Vul & Pashler, 2008). Could reducing the strength of these dependency processes result in errors that are more independent, and therefore result in higher error cancellation rates? For example, if the dermatologist wrongly assessed the first cue as present, but does not remember her initial assessment by the time of the second cue assessment, then this could lead to a more independent, perhaps correct, second cue assessment. The tally (i.e., 1) would still be wrong, but the resulting diagnosis would nevertheless be correct.

We will investigate whether temporally isolating, rather than grouping, the assessments of two cues in the same case reduces the strength of the three dependency processes. In practice this can be realized by introducing separate assessment phases, where in the first phase all cases are assessed according to only the first cue (e.g., asymmetry), and in the second phase all cases are reassessed according to the second cue (e.g., blue/white color). Temporally isolating cue assessments should diminish the strength of the dependency processes because participants may not remember their first assessment for a particular case.

To simplify the experiment, we assumed that a valid strategy is known a priori and used only two cues to assess the final criterion. Furthermore, to control for other factors such as ambiguous cases, untypical appearance, and strength of cues, we created artificial stimuli that resembled skin lesions (Appendix B1).

Apart from the sequential diagnostic procedure, the probabilistic structure of the environment should as well influence a person's decision making process (Todd, Gigerenzer, & ABC Research Group, 2012). For example, in environments with correlated cues, the presence (or absence) of one cue predicts—to the extent of the correlation—the presence (or absence) of the other cue. Individuals will likely learn the underlying cue correlation of the environment and integrate that into their decision-making process. Specifically, we hypothesize that the confirmatory evidence accumulation process is influenced by the underlying correlation between to cues. A high cue correlation implies that congruent cases (i.e., cases where both cues are present or both are absent) occur more often than incongruent cases (i.e., cases where only one cue is present). If individuals search for patterns that are congruent with the hypothesis of the first cue assessment then the confirmatory evidence accumulation process should be amplified in environments where there are more congruent than incongruent cases, resulting in both, an increased discrimination ability for congruent cases, and a decreased discrimination ability for incongruent cases. We therefore implemented two environments: one in which the correlation between the two cues was zero and one in which the correlation was +0.6.

Experiment: What Induces Dependencies and Can They be Reduced?

Stimuli

To simplify our experimental design and analyses we diverged from the three-point checklist of dermoscopy (Zalaudek et al., 2006) to a generic design with synthetically created stimuli that resembled skin lesions, which we called "cell structures." The stimuli were created in Python using the ImaGen package for creating pattern distributions.[1] The stimuli varied on two cues: bright color patches in the cell structure and an irregular network structure containing connected thin lines and cavities (Appendix B, Figure B1). The criterion to be diagnosed was whether the cell structure was "problematic" or "normal." We constructed the criterion value such that two present cues corresponded to a problematic cell structure, and zero or one present cue to a normal cell structure. Participants were asked to assess each cue individually in order to differentiate problematic cell structures from normal cell structures using the tallying rule, which prescribed that a tally of 2 (i.e., a presence of two cues) indicated a problematic cell structure; a lower tally indicated that the stimulus could be regarded as normal. We conducted a pilot experiment and adjusted the strength of the cues in the cell structures until participants' average performance was approximately 70% correct—better than chance, yet leaving room for improvement.

Two Statistical Environments: Zero Versus Positive Intercue Correlation

To investigate whether different correlations between cues influence the dependency processes, we used the ϕ coefficient, a measure of association for two binary variables, to create two cue correlation environments: high correlation (ϕ = +0.6) and zero correlation (ϕ = 0; Table 1).[2] Participants were randomly assigned to one of the environments. Each environment had 250 stimuli and all participants within one environment saw the same set of stimuli. The base rate of the presence of each cue was set to 0.5 for both cues in both environments. Furthermore, with a tallying threshold of 2 to issue a problematic diagnosis, the base rate of the criterion event was 0.25 in the no-correlation environment and 0.4 in the high-correlation environment. The positive predictive value (PPV) per cue for the criterion event was 0.5 in the no-correlation enviroment and 0.8 in high-correlation environment.[3]

Assessment Conditions: Grouped Versus Isolated Assessment of Cues

Staying close to the sequential diagnostic procedure of the three-point checklist (Argenziano et al., 2003; Zalaudek et al., 2006), participants evaluated cases in the grouped cue assessment condition according to

[1] python.org; https://ioam.github.io/imagen/

[2] $\phi = \frac{n_{++}n_{--} - n_{+-}n_{-+}}{\sqrt{n_{+r}n_{-r}n_{c-}n_{c+}}}$, where $n_{++}, n_{+-}, n_{-+}, n_{--}$, are numbers of observations per cell that sum to n, the total number of observations. $n_{+r}n_{-r}$ are the marginal sums of rows, and $n_{c-}n_{c+}$ are the marginal sums of columns (Table 1).

[3] $PPV = \frac{\text{number of true positives}}{\text{number of true positives + number of false positives}}$, where a "true positive" is the event that for example cue S was present and the objective diagnosis was problematic, and a "false positive" is the event that for example cue S was present and the objective diagnosis was normal.

Table 1. Frequency table of structure (S) and color (C) cues in both correlation environments

a) $\phi = 0$

	S+	S -
C+	63	62
C -	62	63

b) $\phi = 0.6$

	S+	S -
C+	100	25
C -	25	100

the first cue, then the second cue, and finally offered an overall diagnosis. That is, participants made three subdecisions before assessing the consecutive case (Figure 1).

Conversely, in the isolated cue assessment condition, participants first evaluated all cases according to one cue (e.g., bright color patches; phase 1). They then saw the same ordered set of cases again assessed the presence of the second cue (e.g., irregular network structure) and gave their final diagnosis (phase 2). Thus, participants in the isolated condition saw all cases twice, whereas participants in the grouped condition saw the cases only once (Figure 1).

Procedure

The experiment was conducted at the laboratory of the Center for Adaptive Rationality at the Max Planck Institute for Human Development in Berlin. At the beginning of the experiment, we collected participants' age, gender, field of study (medicine or other), and experience with dermoscopy (the examination of skin

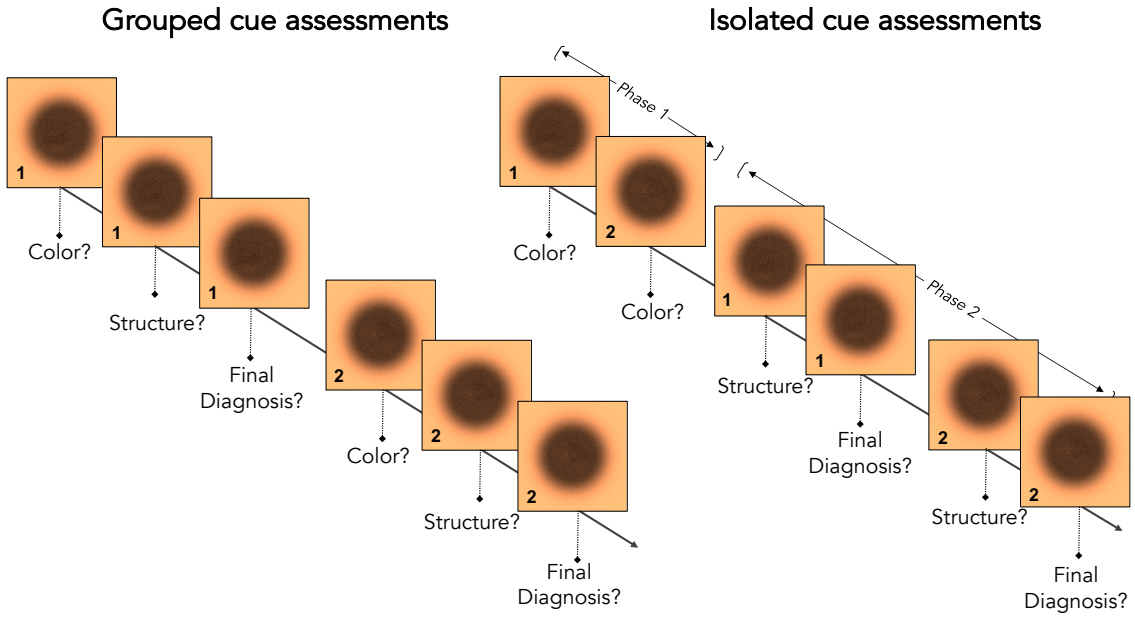


Figure 1. Cue assessment conditions. Numbers indicate individual cases. In the grouped condition (left) a case (1) is fully assessed (i.e., color, structure, final diagnosis) before the next case (2) is assessed. In the isolated condition (right), all cases (1–2) are first assessed according to one cue (color, Phase 1), then all cases (1–2) are assessed according to the second cue (structure) and a final diagnosis is made (Phase 2).

lesions with a dermatoscope).[4] The experiment procedure was structured into three parts—tutorial, training, and testing—and consisted of 14 blocks in total, with 25 cases per block. The first two blocks were part of the training phase and the remaining blocks 3–14 belonged to the testing phase. In total each participant assessed 250 cases.

Tutorial. In the tutorial we informed our participants that the task was similar to a diagnostic procedure for classifying skin lesions in professional practice, but that the stimuli in the experiment were synthetically created cell structures resembling skin lesions. Participants were introduced to the two-point checklist and saw in total twelve example cases of cell structures in the following order: three containing the first cue (e.g., bright color patches), three cell structures without cues, three containing the second cue (i.e., irregular network structure), and again three cell structures without cues. Participants were also provided with a handout summarizing the information from the tutorial and were instructed to study the handout before continuing with one practice case. By completing the practice case participants were familiarized with the actual task: They saw the task structure, pressed the response keys, and could only continue to the next part once they pressed the correct keys. In total participants made three subdecisions per case—first cue, second cue, and final diagnosis—before assessing the consecutive case. For each subdecision participants provided a confidence judgment. Furthermore, we recorded reaction times for decisions and confidence judgments. Participants first evaluated the presence of one cue (e.g., irregular network structure), and provided their confidence in their decision on a half range [50%, 60%, 70%, ..., 100%] probability scale. They then evaluated the presence of the second cue (e.g., bright color patches), followed by a confidence judgment in their decision. Finally, they applied the tallying rule to make an overall diagnosis ("problematic" vs. "normal"), again followed by a confidence judgment in their overall diagnosis. We have not yet analyzed confidence judgments.

Training. In the two training blocks participants were provided with immediate feedback after each subdecision. During the first block, participants could review the handout, which explained the two-point checklist and the visual appearance of the cues. After the first training block, the handout was taken away. All participants were trained in the grouped cue assessment format, but experienced both assessment conditions in separate phases in the testing phase.

Testing. In the testing phase (blocks 3–14) participants were no longer provided with feedback. During the whole experiment participants were free to take as much time as needed to make a decision, yet were informed that in order to finish within approximately one hour, they should not take more than 10 seconds per case.

Participants

100 participants (54 female, median age = 27) were recruited from the subject pool of the Center for Adaptive Rationality at the Max Planck Institute for Human Development. Participants received a show-up fee of €14

[4] Nine participants studied medicine, and three had experience with dermoscopy. In the analyses presented here, we did not take field of study or experience into account.

and were additionally rewarded with €0.01 per correct subdecision; in total participants could earn up to €21.50.

Assignment to environments, conditions and cue order. Participants were randomly assigned to either the zero or 0.6 correlation environment. The order of assessment conditions was counterbalanced across participants, such that half of the participants started the testing phase in the grouped condition, the other half in the isolated condition. Likewise, the order of cues was counterbalanced across participants, such that half of participants—in both assessment conditions—evaluated first the color cue and then the structure cue, and the other half assessed first structure then color.

Statistical Analysis

We modeled our data using a signal detection theory (SDT) framework (Macmillan & Creelman, 2004). SDT classifies binary decisions into four categories: hits, misses, false alarms, and correct rejections (Table 2). A hit corresponds to a "yes" response for a signal trial (i.e., a trial with a truly present cue), while a miss corresponds to a "no" response for a signal trial. A false alarm implies a "yes" response for a noise trial (i.e., a trial where the cue was absent) and a correct rejection means a "no" response for a noise trial.

The basic idea behind SDT is that signal and noise trials can be quantified as values on an arbitrary unidimensional "strength" scale. Whenever individuals see either a signal or noise trial, they experience them as intensities that vary according to a unimodal distribution along this strength scale. If a person can discriminate between signal and noise trials, then in SDT terms the signal distribution has, on average, higher strength values than the noise distribution. SDT assumes that a person has an a priori fixed criterion value which is compared to the strength of a trial. If the strength of the trial exceeds the criterion value, a "yes" response results; if not, the result is a "no" response. In the standard, Gaussian equal variance SDT model, variances of both distributions are set to one and the mean of the noise distribution is set to zero. The mean of the signal distribution is d', which makes d' a measure of discrimination, because it corresponds to the distance between the means of the noise and signal distributions. An increasingly higher d' indicates an increasingly higher discrimination ability. When signal and noise trials occur equally often, $d'/2$ constitutes the unbiased criterion value. The distance between a person's actual criterion and the unbiased criterion is denoted c, which makes c a measure of bias. Positive values of c indicate a strict threshold, that is, a bias towards saying no, whereas negative values of c indicate a lenient threshold, that is, a bias towards saying yes (Macmillan & Creelman, 2004).

Table 2. Signal Detection Theory Categories

	Signal trial	**Noise trial**
Yes response	Hit	False alarm
No response	Miss	Correct rejection

To test our hypotheses we implemented a hierarchical SDT model and used Bayesian estimation techniques (Kruschke, 2014; Lee & Wagenmakers, 2014) to estimate the model parameters and the effects of the assessment conditions and environments on those parameters (Appendix B, section B2).[5] The hierarchical structure implies that each participant-level parameter (c and d') of the SDT model comes from a higher order group-level distribution. We described the higher order group-level distributions for c and d' with a Gaussian normal distribution,

$$\begin{aligned} c &\sim N(\mu, \tau) \\ d' &\sim N(\mu, \tau), \end{aligned} \tag{1}$$

where parameters μ and τ are the mean and precision (the inverse of variance).[6] The resulting posterior distributions of the parameters illustrate the credibility given the data. We summarize the posterior distributions by reporting medians as point estimates and the 95% credible interval (CI)—that is, the upper and lower values between which 95% of the samples fall. When displaying effect sizes in figures, we highlight a "region of practical equivalence" for which Cohen's d's effect size is conventionally considered to be small (from -0.1 to +0.1; Kruschke, 2013). Analyses were conducted in the statistical computing software R (R Core Team, 2013) using JAGS (Plummer et al., 2003) via the R2jags package (Su & Yajima, 2015).

Hypotheses

We preregistered our hypotheses before data collection at https://aspredicted.org/blind.php?x=jh9ad7.

Hypothesis 1: Confirmatory response effect

1. If the first cue was judged as absent, criterion c for the second cue assessment will be more strict (i.e., the participant is more likely to assess the second cue as absent) than when the first cue is judged as present.

2. In the isolated condition (compared to the grouped condition), criterion c for the second cue assessment should depend less on whether the first cue was judged as present or absent.

Hypothesis 2: Confirmatory evidence accumulation

1. If the first cue assessment is congruent with the true cue value of the second cue, discrimination d' for second cue assessments will be higher than when the first cue assessment is incongruent with the true

[5] Given that we collected response times for each decision, it would have been possible to analyze our data with a hierarchical drift diffusion model (DDM; Ratcliff & McKoon, 2008; Vandekerckhove, Tuerlinckx, & Lee, 2011). Drift diffusion models can quantitatively predict both choice and response times, whereas SDT models predict choices only. This is an advantage of the DDM because it is possible that some dependency processes affect response times but not the observed choices, or multiple, even antagonistic effects on different decision-making parameters (Pleskac, Cesario, & Johnson, 2018). Because of these advantages we began our analyzes with a hierarchical DDM but eventually had to conclude that our data could not be fitted by the DDM as it cannot account for the multimodal response time distributions we observed (Figure B2 in Appendix B)—rendering the parameter estimates of the model invalid for interpretation.

[6] For priors for those Gaussian group-level distributions, we used a Gaussian normal distribution $N(0, .001)$ for μ and a gamma distribution $G(.001, .001)$ for τ (precision).

cue value of the second cue. This effect should be stronger when $\phi = 0.6$, because congruent cases occur more often, than when $\phi = 0$.

2. In the isolated condition (compared to the grouped condition), discrimination d' for second cue assessments should depend less on whether the first cue assessment was congruent or incongruent with the true cue value of the second cue.

Hypothesis 3: Relevance effort effect

1. If the first cue was judged as present (relevant second cue assessment), discrimination d' of second cue assessments should be higher than when the first cue was judged as absent (irrelevant second cue assessment).

2. In the isolated condition (compared to the grouped condition), discrimination d' for second cue assessments should depend less on whether the first cue was judged as present or absent.

Results

Descriptive Results of Training and Testing Phase

Training phase. Figure 2 illustrates participants' balanced accuracy during the training phase.[7] Participants categorized the two cues and provided final diagnoses with median balanced accuracy ranging between .68 and .74 in the $\phi = 0$ environment and between .69 and .79 in the $\phi = 0.6$ environment.

Testing phase Participants' median balanced accuracies improved from the training to the testing phase. In general, first cue assessments had higher median balanced accuracies (ranging between .80 and .87) than second cue assessments (ranging from .73 to .76) throughout environments and conditions (Figure 3). We did not expect this finding and will return to this in the General Discussion. The next sections present the effects of the dependency processes in grouped and isolated cue assessments, and the differences between both conditions. Finally, we outline exploratory findings about the downstream consequences of the tallying rule on the final diagnosis.

Dependency in Grouped Cue Assessments

We hypothesized that the sequential diagnostic procedure of tallying (i.e., grouped cue assessments) would induce dependencies between cue assessments. The confirmatory response effect hypothesis (Hypothesis 1) predicted an effect on the criterion c, a measure of bias. Recall that positive values of c indicate a strict threshold, that is, a bias towards saying "absent," and negative values of c indicate a lenient threshold, that is, a bias towards saying "present." The confirmatory response effect hypothesis predicted a positive c (i.e.,

[7] Balanced accuracy = (hit rate + 1 - false alarm rate)/2. Balanced accuracy can deviate from proportion correct whenever the number of signal and noise trials differ. The high correlation environment had more signal than noise trials.

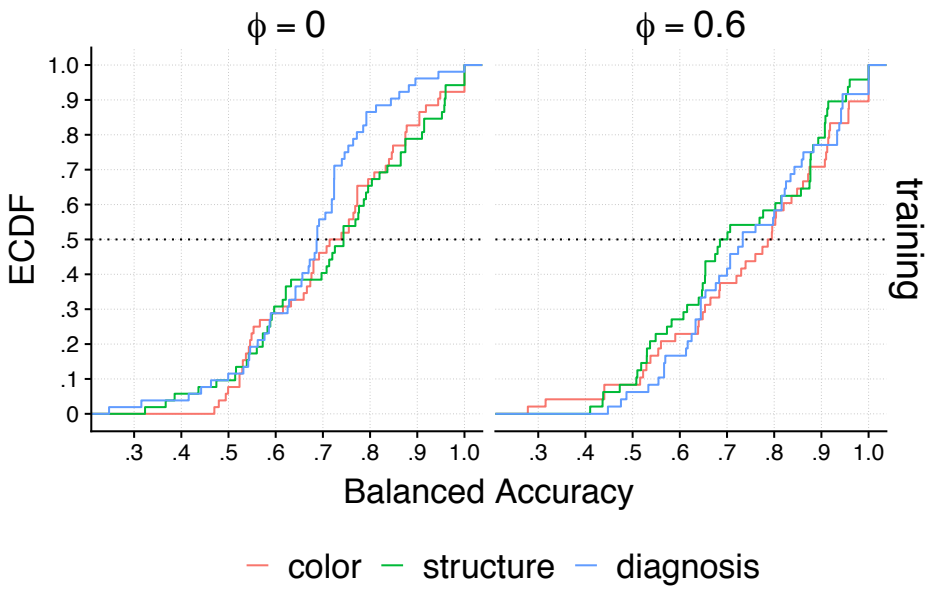


Figure 2. Training results. Empirical cumulative distribution function (ECDF; y-axis) of participants' balanced accuracies (x-axis) in the training phase, separately per environment and cue assessment. The ECDF depicts the proportion of observations with a value at or below the value on the x-axis. The dotted horizontal line (at 50%) shows the median. Median values ranged between .68 and .74 when $\phi = 0$ and between .69 and .79 when $\phi = 0.6$.

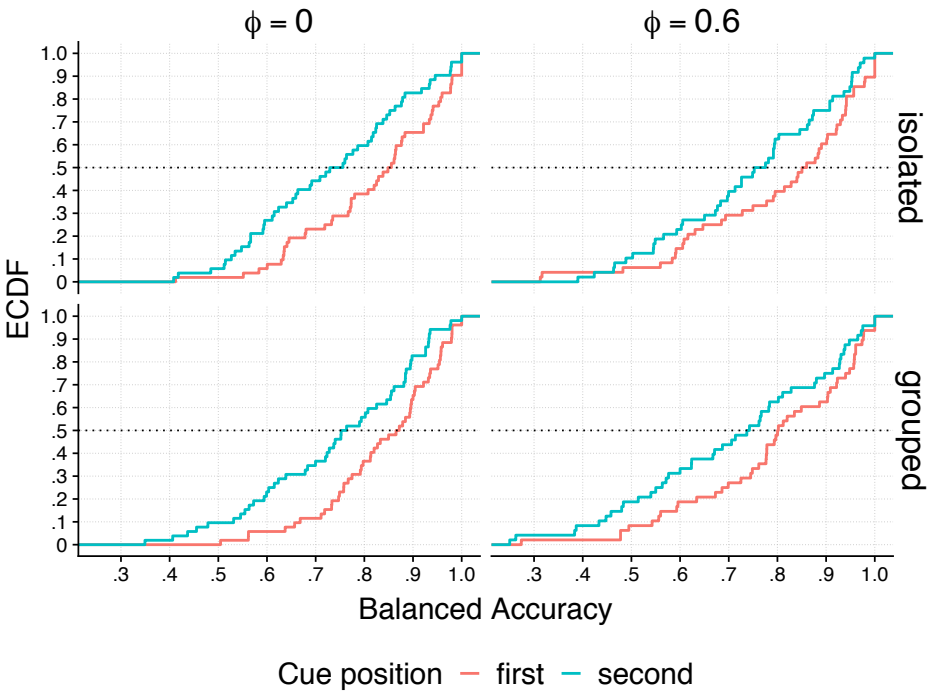


Figure 3. Testing results. Empirical cumulative distribution function (ECDF; y-axis) of participants' balanced accuracies (x-axis) in the testing phase, separately per environment, assessment condition, and cue position. The ECDF depicts the proportion of observations with a value at or below the value on the x-axis. The dotted horizontal line (at 50%) shows the median. Median values for first cue assessments (red line) ranged between .80 and .86, and between .74 and. 76 for second cue assessments (blue line).

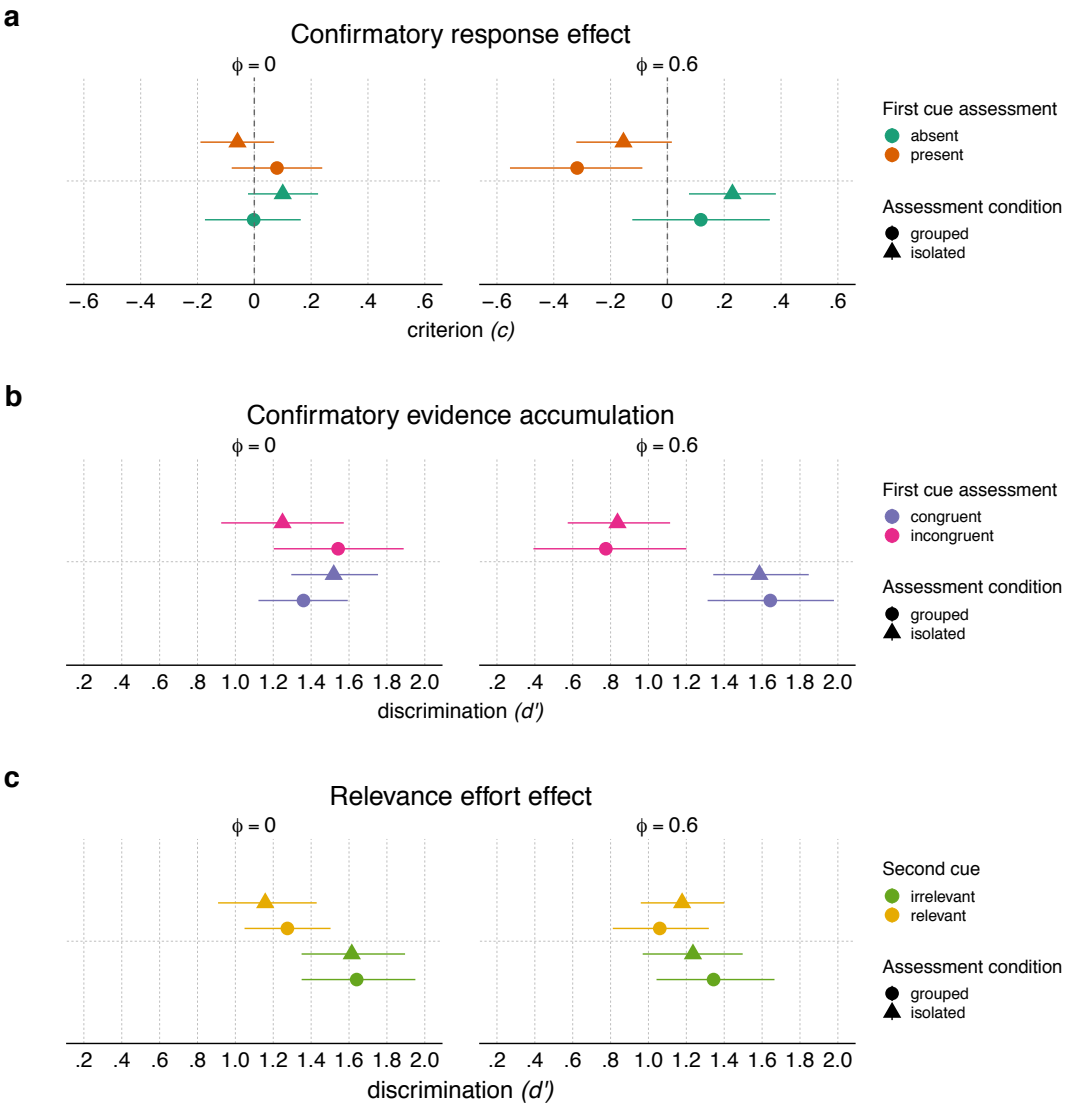


Figure 4. Posterior estimates of signal detection measures (criterion c and discrimination d') for second cue assessments at the group level. Symbols indicate the median posterior value, bars the 95% credible interval (CI), separately per environment, hypothesis, and assessment condition. For the criterion c positive values indicate a bias towards responding "absent," while negative values indicate a bias towards responding "present." Discrimination d' is a measure of participants' ability to discriminate between the signal (i.e., present) and noise (i.e., absent) trials and corresponds to the distance between signal and noise distributions. Increasingly positive values indicate increasing discrimination. Differences are stronger in the $\phi = 0.6$ environment than in the $\phi = 0$ environment. a) Estimates in the $\phi = 0.6$ environment were in line with the predicted pattern of the confirmatory response effect hypothesis (Hypothesis 1). When the first cue was assessed as absent, criterion c for second cue assessments was higher (i.e., participants were biased to judge the second cue assessment to be absent) than when the first cue was assessed as present (i.e., participants were biased to say present for the second cue assessment). b) Estimates in the $\phi = 0.6$ environment corresponded with the predicted pattern of the confirmatory evidence accumulation hypothesis (Hypothesis 2). When the first cue assessment was congruent with the true state of the second cue, discrimination d' was higher than when the first cue assessment was incongruent with the true state of the second cue. c) Estimates differed more in the $\phi = 0$ environment, but in the opposite direction of what the relevance effort effect hypothesis (Hypothesis 3) predicts: When the second cue assessment was irrelevant for the overall diagnosis, discrimination d' was higher than when it was relevant for the overall diagnosis.

biased towards an "absent" assessment) for second cue assessments, when the first cue was judged as absent and a negative c (i.e., biased towards a "present" assessment) when the first cue was judged as present in the grouped cue assessment condition. Figure 4 summarizes the posterior distributions for the signal detection measures at the group level (i.e., criterion c and discrimination d') and Figure 5 summarizes the posterior distributions of the differences between assessment conditions (expressed as Cohen's d effect sizes). In line with this hypothesis, we observed a confirmatory response effect for grouped cue assessments in the $\phi = 0.6$ environment (Cohen's d = 0.55, 95% − CI [0.12, 0.99]). This result shows that second cue assessments are biased towards the initial cue assessment (Figure 4a & Figure 5a). We did not find this effect in the $\phi = 0$ environment (Cohen's d = -0.15, 95% − CI [-0.58, 0.26]; Figure 4a, Table 3), suggesting that the underlying cue correlation moderates the confirmatory response effect.

The confirmatory evidence accumulation hypothesis (Hypothesis 2) predicted an effect on the discrimination measure d', a measure of how well one can differentiate between signal (i.e., "present") and noise stimuli (i.e., "absent"). Increasing values of d' correspond to increasingly higher discrimination ability. The hypothesis predicts a higher d' for second cue assessments if the first cue assessment was congruent with the true state of the second cue compared to when the first cue assessment was incongruent with the true state of the second cue. We also expected that this effect would be stronger when $\phi = 0.6$ than when $\phi = 0$. Consistent with our hypothesis, we observed a confirmatory evidence accumulation effect for grouped cue assessments in the $\phi = 0.6$ environment (Cohen's d = 0.73, 95% − CI [0.29, 1.22]; Table 3), showing that discrimination d' for second cue assessments was better (i.e., higher) when the first cue assessment was congruent with the true state of the second cue (Figure 4b & Figure 5b). We did not find this effect in the $\phi = 0$ environment (Cohen's d = -0.19, 95% − CI [-0.64, 0.25]; Table 3), suggesting that the underlying cue correlation moderated the effect of confirmatory evidence accumulation (Figure 4b, Table 3).

Finally, according to the relevance effort effect hypothesis (Hypothesis 3), d' for second cue assessments in the grouped condition was expected to be higher (i.e., better discrimination) when the second cue was relevant for the overall diagnosis, compared to when it was not relevant. With a tallying threshold of 2, the second cue was relevant for the overall diagnosis whenever the first cue was assessed as present and irrelevant otherwise. In contrast to our hypothesis, the results showed the reverse pattern when $\phi = 0$ (Cohen's d = -0.44, 95% − CI [-0.87, 0.01], Figure 5c), that is, a higher discriminability when the second cue was irrelevant for the overall diagnosis. Differences in the $\phi = 0.6$ environment were less pronounced (Cohen's d = -0.35, 95% − CI [-0.83, 0.13]; Figure 5c). However, in both environments the posterior distributions' 95% − CI's overlapped with the region of practical equivalence (Table 3).

Isolating Cue Assessments

The purpose of the isolated cue assessment condition was to investigate whether the hypothesized dependency effects could be reduced. In this condition all cases were first assessed according to one cue, then all cases were again assessed according to the second cue and the final diagnosis. Figure 5 shows the dependency effects

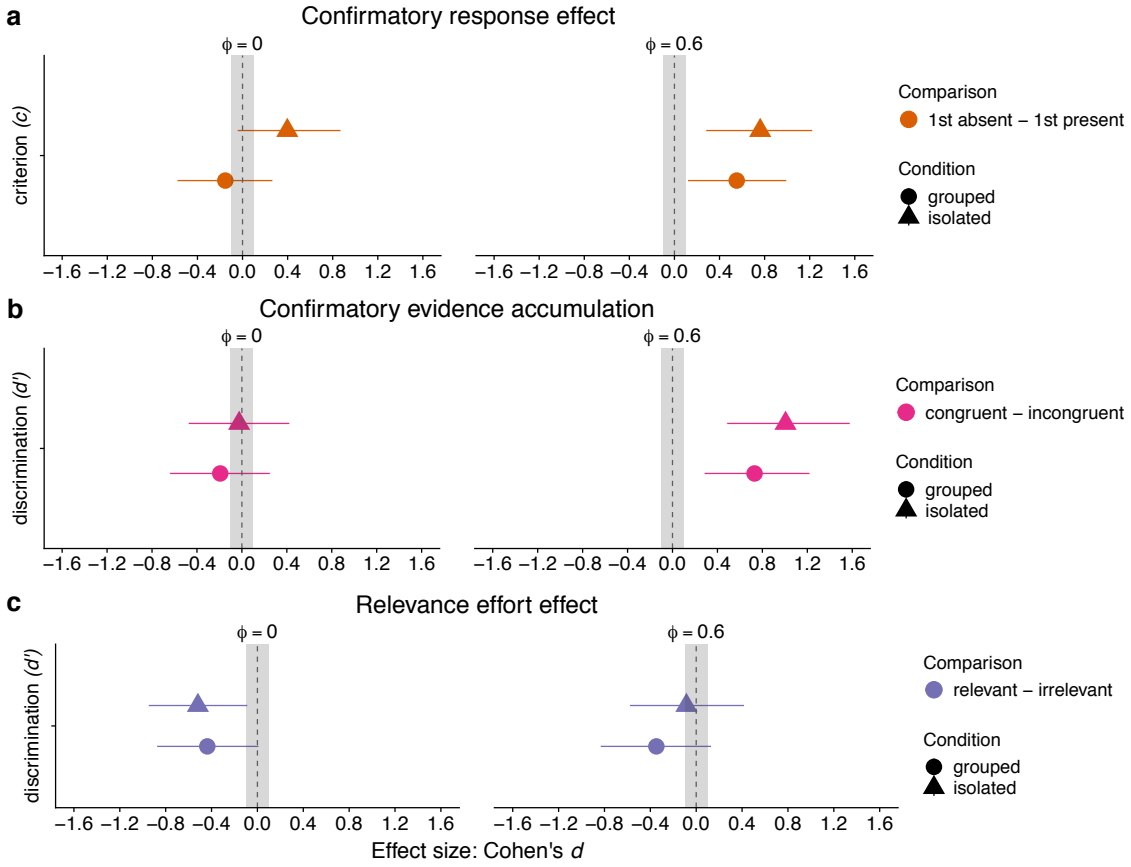


Figure 5. Dependency effects on second cue assessments as revealed by signal detection measures (criterion c and discrimination d') at the group level. The x-axis shows the posterior distributions of the respective differences at the group level (expressed as Cohen's d effect sizes). Symbols indicate the median posterior value, bars the 95% credible interval (CI). The shaded region ranging between −0.1 and +0.1 highlights the region of practical equivalence, for which Cohen's d effect size is conventionally considered to be small (from −0.1 to +0.1). Comparisons are between second cue assessments, depending on whether the first cue was a) assessed as absent/present, b) congruent/incongruent with the true state of the second cue, or c) whether the second cue assessment was relevant/irrelevant for the final diagnosis. Results are shown separately per environment, hypothesis, and assessment condition. We predicted positive effect sizes for the grouped condition and smaller (or even zero) effect sizes in the isolated condition. In a) and b) but not in c) grouped cue assessments showed the predicted effects when $\phi = 0.6$, but not when $\phi = 0$. Contrary to predictions, isolated cue assessments showed positive effect sizes in a) in both environments and in b) for $\phi = 0.6$, suggesting that dependency effects could not be reduced. c) We observed the opposite of our predicted pattern for both grouped and isolated cue assessment when $\phi = 0$, but not when $\phi = 0.6$.

between cue assessments for the isolated condition. We observed a confirmatory response effect (Figure 5a; Cohen's $d = 0.76$, $95\% - CI$ [0.28, 1.22]) and a confirmatory evidence accumulation effect (Cohen's $d = 1.01$, $95\% - CI$ [0.49, 1.58]) in the $\phi = 0.6$ environment (Figure 5b), but no relevance effort effect (Cohen's d = -0.08, $95\% - CI$ [-0.58, 0.42]) in the $\phi = 0.6$ environment. All effect sizes for the hypotheses in the isolated cue assessment condition can be found in Table 3.

Figure 6 and Table 3 illustrate the effect sizes of the differences between grouped and isolated cue assessments. Results show that almost all effect sizes are around zero (or include the zero value) for each hypothesized dependency effect and in both environments, implying that isolating cue assessments in time did not reduce dependency between cue assessments (Figure 6, Table 3).

Table 3. Summary of effects within and between assessment conditions

Variable	Cohen's d	$95\% - CI$
A) Dependency effects in grouped cue assessments		
Environment: $\phi = 0$		
Confirmatory response effect (c)	-0.15	[-0.58, 0.26]
Confirmatory evidence accumulation (d')	-0.19	[-0.64, 0.25]
Relevance effort effect (d')	-0.44	[-0.87, 0.01]
Environment: $\phi = 0.6$		
Confirmatory response effect (c)	0.55	[0.12, 0.99]
Confirmatory evidence accumulation (d')	0.73	[0.28, 1.22]
Relevance effort effect (d')	-0.35	[-0.83, 0.13]
B) Dependency effects in isolated cue assessments		
Environment: $\phi = 0$		
Confirmatory response effect (c)	0.40	[-0.04, 0.87]
Confirmatory evidence accumulation (d')	-0.02	[-0.47, 0.42]
Relevance effort effect (d')	-0.52	[-0.95, -0.09]
Environment: $\phi = 0.6$		
Confirmatory response effect (c)	0.76	[0.28, 1.22]
Confirmatory evidence accumulation (d')	1.01	[0.49, 1.58]
Relevance effort effect (d')	-0.08	[-0.58, 0.42]
C) Assessment effects: Grouped minus isolated condition		
Environment: $\phi = 0$		
Confirmatory response effect (c)	-0.55	[-1.19, 0.07]
Confirmatory evidence accumulation (d')	-0.17	[-0.52, 0.18]
Relevance effort effect (d')	0.08	[-0.53, 0.69]
Environment: $\phi = 0.6$		
Confirmatory response effect (c)	-0.21	[-0.84, 0.44]
Confirmatory evidence accumulation (d')	-0.27	[-1.01, 0.43]
Relevance effort effect (d')	-0.26	[-0.95, 0.42]

Note: $95\% - CI = 95\%$ credible interval. Effect sizes (Cohen's d) are calculated so that positive values correspond to effects in the predicted direction.

Exploratory Analysis

Dependencies in sequential assessments will not only affect the accuracy of cue assessments, but also that of final diagnoses. Here we present a descriptive exploratory analysis on the *given* and *implied* diagnoses. A given diagnosis is simply a participant's answer to the question, "Is this cell structure problematic?" An implied diagnosis is the diagnosis that would follow from strictly applying the tallying rule to participants' cue assessments. A given diagnosis can differ from an implied diagnosis in both the grouped assessment and the isolated assessment conditions. In the grouped assessment condition this could occur if a participant does not strictly apply the tallying rule, perhaps due to having changed their mind about one cue assessment and therefore "internally" changing the tally and hence the diagnosis. In the isolated condition a participant's

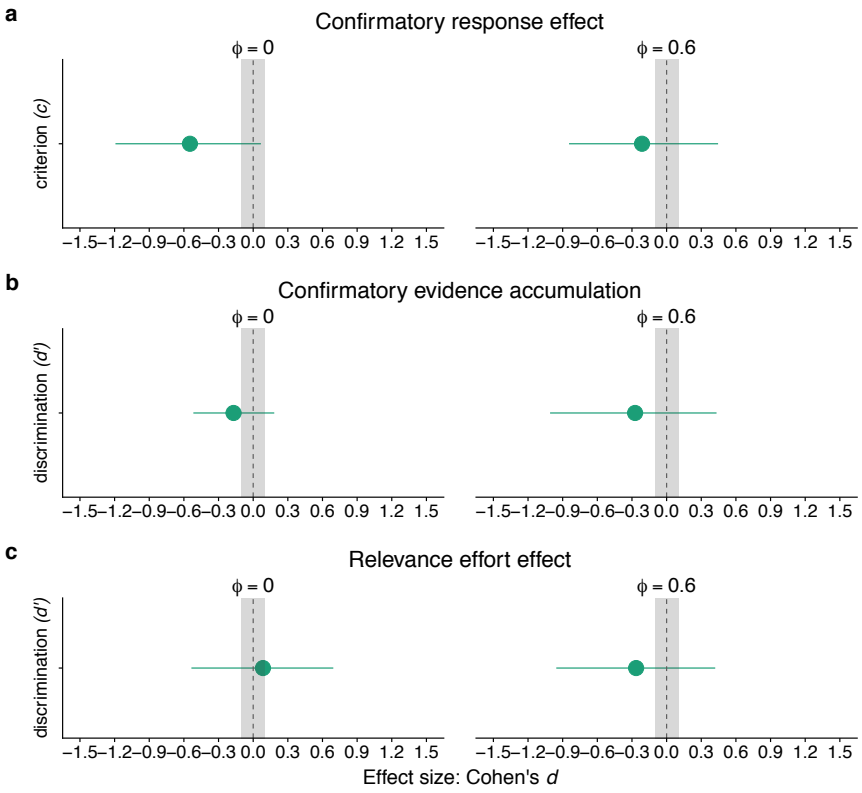


Figure 6. Difference between grouped versus isolated cue assessment on signal detection measures c and d'. The x-axis shows the posterior distributions of differences (expressed as Cohen's d effect sizes). Symbols indicate the median posterior value, bars the 95% credible interval (CI). The shaded region ranging between −0.1 and +0.1 highlights the region of practical equivalence, for which Cohen's d effect size is conventionally considered to be small (from −0.1 to +0.1). Results are shown separately per environment and hypothesis. We predicted positive effect size differences (grouped assessment condition minus isolated assessment condition). Differences are either zero or negative, but in all cases effect sizes include the zero value.

implied diagnosis can differ from their given diagnosis if they do not remember their first given cue assessment when they later assesses the second cue; in this scenario, a participant could arrive at a different diagnosis (in principle, this reasoning could apply for the grouped condition as well, but is unlikely, since the two assessments are made successively). Figure 7 suggests that in the isolated condition participants' implied diagnoses were more accurate than their given diagnoses, although the effect sizes were small, with Cohen's $d = 0.27$ in the $\phi = 0.6$ environment and Cohen's $d = 0.24$ in the $\phi = 0$ environment. The implied diagnoses in the isolated condition were better than the given diagnoses of the grouped condition in the $\phi = 0.6$ environment (Cohen's $d = 0.36$), but we observed the reverse pattern in the $\phi = 0$ environment (Cohen's $d = -0.2$).

General Discussion

Each year an estimated 200,000 patients in the United States alone die from preventable medical errors (Andel, Davidow, Hollander, & Moreno, 2012); many more undergo serious harm, disability, and false treatment (Berner & Graber, 2008; Blendon et al., 2002) and suffer the legal and financial consequences thereof (Andel

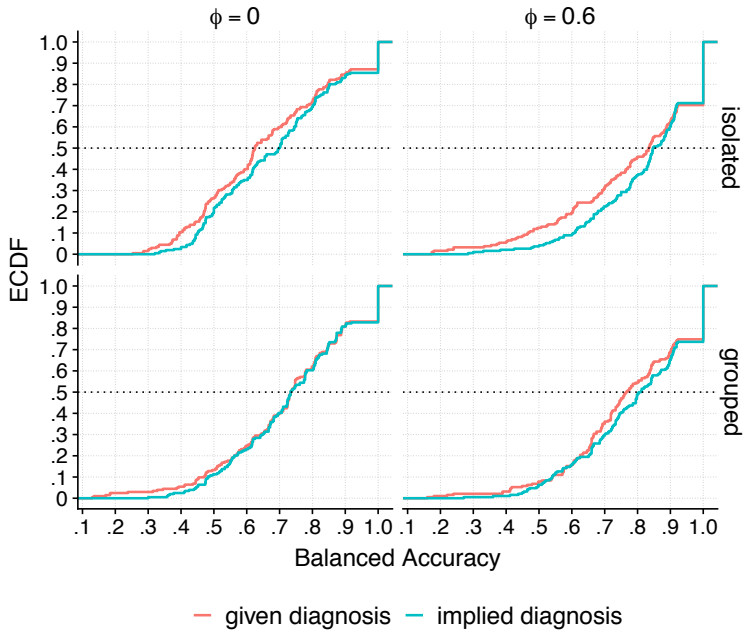


Figure 7. Accuracy of given and implied diagnoses. Empirical cumulative distribution function (ECDF; y-axis) of participants' balanced accuracies (x-axis) for final and implied diagnoses, separately per environment and assessment condition. Implied diagnoses (blue line) are more accurate than the given diagnoses (red line). This trend was stronger in the $\phi = 0.6$ environment and in the isolated assessment condition. The ECDF depicts the proportion of observations with a value at or below the value on the x-axis. The dotted horizontal line (at 50%) shows the median.

et al., 2012; Kahneman, Rosenfield, Gandhi, & Blaser, 2016). Growing literature suggests that a majority of diagnostic errors can be attributed to an individual doctor's cognitive processes (Graber, Franklin, & Gordon, 2005; Hussain & Oestreicher, 2017; Norman & Eva, 2010; but see Sherbino et al., 2012). Identifying the cognitive processes underlying diagnostic errors is thus a major step toward improving health care and patient safety (Thammasitboon & Cutrer, 2013). Here we investigated whether the sequential diagnostic procedure of a tallying rule induces cognitive dependency processes and studied how such dependency processes could be reduced. Based on previous research we enlisted three mutually compatible processes: the confirmatory response effect, which biases an individual towards the hypothesis corresponding to the first piece of evidence they encounter; the confirmatory evidence accumulation process, which influences an individual's evidence accumulation towards the hypothesis corresponding to the first piece of evidence they encounter; and the relevance effort effect, which influences an individual's effort depending on the relevance of the consecutive cues for the final diagnosis. We created two environments with different cue correlations (i.e., $\phi = 0$ and $\phi = 0.6$) and studied whether dependency can be reduced by temporally isolating cue assessments in the same case.

Dependency Processes in Grouped Cue Assessments

Our results showed the presence of the confirmatory response effect and the confirmatory evidence accumulation process in an environment with a positive correlation between cues, and not in an environment where cues

were uncorrelated. However, if the sequential diagnostic procedure of tallying induces cognitive dependency processes, then these processes should also have appeared in the no-correlation ($\phi = 0$) environment. Since we found the effects only in the correlation ($\phi = 0.6$) environment, this suggests that the correlation among cues moderates the dependency processes. Indeed, research has shown that individuals are sensitive to the underlying probabilistic structure of the environment and adapt their decision-making strategies accordingly (Jarecki, Meder, & Nelson, 2018; Pachur & Olsson, 2012; Pleskac & Hertwig, 2014). This raises a normative question that should be addressed in future studies: To what extent it is adaptive to integrate the underlying probabilistic structure of an environment into one's decision-making process? With regard to our confirmatory response effect hypothesis, we speculate that, when making decisions under uncertainty, it may be beneficial for the accuracy of second cue assessments to adapt one's criterion c (i.e., bias) according to the probabilistic environment. For instance, when the presence of the first cue predicts the presence of the second cue (correlated environment) and the first cue is assessed as present, lowering one's threshold for the second cue (i.e., one's bias toward an assessment of "present") could increase the accuracy of the second cue assessment. Conversely, when the absence of the first cue predicts the absence of the second cue and the first cue was assessed as absent, then raising one's threshold for the second cue (i.e., one's bias toward an assessment of "absent") could increase the accuracy of the second cue assessment.

However, to normatively assess the extent to which individuals should adapt their criterion value c not only depends on the consequences for the accuracy of the second cue assessment, but also on the consequences for the overall diagnosis as implied by the tallying rule. Consider a case where the first cue assessment was wrongly judged as present, making an individual more likely to judge the second cue assessment as present. Then, even if this tendency would make it more likely that the second cue is assessed correctly, the downstream consequences of both cues being judged as present can nevertheless lead to a wrong final diagnosis. The downstream consequences of the two cue assessments on the final diagnosis should be addressed in future studies.

Concerning our confirmatory evidence accumulation hypothesis, we predicted that individuals search for patterns that are congruent with their assessment of the first cue, resulting in increased discrimination for congruent cases and decreased discrimination for incongruent cases. This process should be amplified in environments where there are more congruent than incongruent cases (i.e., a high cue-correlation environment). The pattern of our results was in line with the predicted pattern—there was a stronger effect in the correlation compared to the no-correlation environment. It could be that when individuals learn the cue correlation and use the presence of one cue to make inferences about the presence of another cue, this not only affects the response tendency (i.e., criterion c) but also directs an individual's attention more strongly to search for congruent information.

To investigate whether individuals are well adapted to the environment, future research should study whether individuals who displayed a confirmatory response effect also followed a confirmatory evidence accumulation process and vice versa. If both processes happen simultaneously within an individual, how does that affect their accuracy for second cue assessments and the final diagnosis, in comparison to individuals who

only display one or none of the processes? For example, if both processes happen simultaneously, would they cancel or amplify effects on accuracy for second cue assessments and the final diagnosis?

With respect to the role of adaptive decision making in our relevance effort effect hypothesis, there is little reason to assume that adapting one's effort according to the underlying cue correlation in the environment is a reasonable strategy. Rather, it would be desirable to keep one's effort high at any point in time, independent of the cue correlation, in order to maximize the accuracy of second cue assessments.

The finding that we observed the opposite of the expected pattern for the relevance effort effect in the $\phi = 0$ environment remains surprising. We hypothesized that the effort spent on the second cue assessment depends on the relevance of the second cue for the overall diagnosis. If the first cue was assessed as present and therefore the second cue assessment can change the final diagnosis (if the second cue is assessed as present the overall diagnosis is "problematic" and if assessed as absent the final diagnosis is "normal"), the effort in assessing the second cue should stay high. However, when the first cue was assessed as absent, making the second cue irrelevant for the final diagnosis, the effort in assessing the second cue should decline. To our surprise, we observed the opposite effect: Individuals put more effort in their second cue assessment when the second cue was irrelevant for the final diagnosis as compared to when it was relevant. We currently cannot offer an explanation for this result.

Isolating Cue Assessments

We investigated whether temporally isolating cue assessments can reduce the cognitive dependency processes. When assessing the second cue, individuals might not remember their first cue assessment; therefore isolating cue assessments should reduce dependency effects. However, we found no such effects. All dependency processes we found also appeared in the isolated condition and, in fact, the confirmatory response effect appeared in the isolated condition but not in the grouped condition (i.e., for criterion c when $\phi = 0$). This finding implies that participants' decision processes were similar in both assessment conditions, that is, during the second phase of the isolated condition participants were probably reassessing the first cue and only then the second cue. It is possible that participants proceduralized their decision-making behavior in the grouped assessment condition during the training phase, such that they automatically assessed the first cue and then the second cue in the isolated assessment condition. Furthermore, participants were asked to make a final diagnosis in the second phase of the isolated condition, therefore, to give a final diagnosis, they had to take the first cue into account. Because our goal was to investigate whether it is possible to apply the tallying rule but reduce dependency by isolating cue assessments, we deliberately decided to collect a final diagnosis in the isolated condition in order to emphasize the tallying rule—the ultimate goal of the rule is to make a final diagnosis and all hypothesized dependency effects depend on the tally. However, this design may have come at the cost of rendering the isolated condition ineffective.

Future research should address different methods for reducing dependency and investigate whether more independent judgments eventually result in improved final diagnoses. One method could be to randomize

the cue order within participants instead of the current design of keeping it constant. On the one hand, randomizing the cue order might prevent individuals from proceduralizing the sequential assessment of both cues and therefore reduce the dependency processes. On the other hand, randomizing the order may make assessing the cues more effortful—and potentially error-prone—because of the mental costs of task-switching. Another possibility would be to not ask for a final diagnosis in the second assessment phase, but instead have three separate assessment phases: one for cue one, a second for cue two, and a third for the final diagnosis. However, we believe that the best procedure would be to eliminate the final diagnosis altogether and only aggregate the individual cue assessments mechanically using the tallying rule. Depending on the performance of such a procedure, new assessment methods can be developed.

Performance of Second Cue Assessments

Our descriptive analysis showed that throughout the training and testing phase, in both assessment conditions and cue correlation environments, participants' accuracy on second cue assessments was consistently lower than that of their first cue assessments—independent of which of the two cues (i.e., color or structure) was first (Figure 3). This finding is unexpected. Future studies should investigate whether the effect diminishes when individuals merely learn to categorize cues without learning the tallying rule.

Next Steps in Studying Cognitive Dependencies in Sequential Diagnostic Reasoning Tasks

Future research should expand on at least three points. First, we attempted to isolate cue assessments to decrease the dependency between cue assessments, but we might not have fully exploited the potential of isolating cue assessments. As discussed above ("Isolating Cue Assessments") participants might have proceduralized the grouped cue assessment format, leading them to consistently assess the first cue and then the second cue even in the isolated condition. Furthermore, in the second phase of the isolated condition, participants also had to provide a final diagnosis, which required them to reassess the first cue. Possible alternatives to reduce dependency would be to randomize the cue order within participants, or to introduce three assessment phases (i.e., one for cue one, a second for cue two and a third for the final diagnosis). The strongest test, however, would be to omit the final diagnosis altogether and mechanically derive the implied final diagnosis by applying the tallying rule to participants' cue assessments.

Second, another source of noise in cue assessments could result from interference in visual information processing. Anticipating the order of cue assessments (e.g., first cue A, then cue B and the final diagnosis, before moving to the next case), might result in holistic processing of both visual cues and potentially lead to a diminished discrimination ability for both cues. Investigating whether individuals who learned to discriminate only one cue perform differently from individuals who learned to discriminate both could be used to further develop methods to harness the wisdom of crowds, where diagnostic judgments are aggregated across individuals (Kämmer, Hautz, Herzog, Kunina-Habenicht, & Kurvers, 2017; Kurvers et al., 2016). For example, one

method could be to aggregate cue assessments of individuals who were only trained in assessing one of several cues and then mechanically derive the final diagnosis from the individuals' aggregated cue assessments.

Third, in a typical clinical setting physicians have a limited amount of time to dedicate to each patient. In our experiment we investigated how individuals make decisions when there are no time constraints. It would be informative to study whether dependency effects increase or decrease under time pressure. On the one hand, without time pressure individuals can harness their skills to thoroughly analyze a case (Mamede et al., 2010). On the other hand, a skilled person's initial "gut feeling" might be the most accurate choice (Norman et al., 2014; Sherbino et al., 2012) and any further evidence accumulation could be distorted by cognitive processes such as the confirmatory evidence accumulation or the confirmatory response effect.

Diagnostic errors can seriously endanger patients' health. The literature suggests that a majority of errors result from preventable cognitive processes (Andel et al., 2012; Berner & Graber, 2008; Blendon et al., 2002). Our study provides first insights into the cognitive processes during sequential diagnostic decision making, taking into account the probabilistic structure of the environment. Even if individuals exhibit cognitive dependency processes, future research should investigate to what extent they are adaptive and how else diagnostic errors could be prevented.

References

Andel, C., Davidow, S. L., Hollander, M., & Moreno, D. A. (2012). The economics of health care quality and medical errors. *Journal of Health Care Finance*, *39*(1), 39–50.

Argenziano, G., Soyer, H. P., Chimenti, S., Talamini, R., Corona, R., Sera, F., ... Kopf, A. W. (2003). Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the internet. *Journal of the American Academy of Dermatology*, *48*(5), 679–693. doi: 10.1067/mjd.2003.281

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178. doi: 10.1146/annurev.psych.56.091103.070217

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5), S2–S23. doi: 10.1016/j.amjmed.2008.01.001

Blendon, R. J., DesRoches, C. M., Brodie, M., Benson, J. M., Rosen, A. B., Schneider, E., ... Steffenson, A. E. (2002). Views of practicing physicians and the public on medical errors. *New England Journal of Medicine*, *347*(24), 1933–1940. doi: 10.1056/NEJMsa022151

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer, P. M. Todd, & ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 97–118). New York, NY: Oxford University Press.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. doi: 10.1037/0003-066X.34.7.571

Germar, M., Albrecht, T., Voss, A., & Mojzisch, A. (2016). Social conformity is due to biased stimulus processing: Electrophysiological and diffusion analyses. *Social Cognitive and Affective Neuroscience*, *11*(9), 1449–1459. doi: 10.1093/scan/nsw050

Germar, M., Schlemmer, A., Krug, K., Voss, A., & Mojzisch, A. (2014). Social influence and perceptual decision making: A diffusion model analysis. *Personality and Social Psychology Bulletin*, *40*(2), 217–231. doi: 10.1177/0146167213508985

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior.* Oxford, United Kingdom: Oxford University Press.

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine*, *165*(13), 1493–1499. doi: 10.1001/archinte.165.13.1493

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237. doi: 10.1111/j.1467-9280.2009.02271.x

Hussain, A., & Oestreicher, J. (2017). Clinical decision-making: Heuristics and cognitive biases for the ophthalmologist. *Survey of Ophthalmology*, 119–124. doi: 10.1016/j.survophthal.2017.08.007

Jarecki, J. B., Meder, B., & Nelson, J. D. (2018). Naïve and robust: Class-conditional independence in human classification learning. *Cognitive Science*, *42*(1), 4–42. doi: 10.1111/cogs.12496

Johnson, S. G. B., Rajeev-Kumar, G., & Keil, F. C. (2016). Sense-making under ignorance. *Cognitive Psychology*, *89*, 39–70. doi: 10.1016/j.cogpsych.2016.06.004

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*, *94*(10), 38–46.

Kämmer, J. E., Hautz, W. E., Herzog, S. M., Kunina-Habenicht, O., & Kurvers, R. H. J. M. (2017). The potential of collective intelligence in emergency medicine: Pooling medical students' independent decisions improves diagnostic performance. *Medical Decision Making*, *37*(6), 715–724. doi: 10.1177/0272989X17696998

Kattah, J. C., Talkad, A. V., Wang, D. Z., Hsieh, Y.-H., & Newman-Toker, D. E. (2009). HINTS to diagnose stroke in the acute vestibular syndrome: Three-step bedside oculomotor examination more sensitive than early mri diffusion-weighted imaging. *Stroke*, *40*(11), 3504–3510. doi: 10.1161/strokeaha.109.551234

Kostopoulou, O., Mousoulis, C., & Delaney, B. (2009). Information search and information distortion in the diagnosis of an ambiguous presentation. *Judgment and Decision Making*, *4*(5), 408–418.

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 267–301). New York, NY: Cambridge University Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi: 10.1037/a0029146

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London, United Kingdom: Academic Press.

Kurvers, R. H. J. M., 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 of the United States of America*, *113*(31), 8777–8782. doi: 0.1073/pnas.1601827113

Lange, N. D., Thomas, R. P., & Davelaar, E. J. (2012). Temporal dynamics of hypothesis generation: The influences of data serial order, data consistency, and elicitation timing. *Frontiers in Psychology*, *3*, 215–231. doi: 10.3389/fpsyg.2012.00215

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers in social psychology: Social judgment and decision making* (p. 227—242). New York, NY: Psychology Press.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, United Kingdom: Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. New York, NY: Psychology Press.

Makary, M. A., & Daniel, M. (2016). Medical error—the third leading cause of death in the US. *Bmj*, *353*, i2139. doi: 10.1136/bmj.i2139

Mamede, S., Schmidt, H. G., Rikers, R. M. J. P., Custers, E. J. F. M., Splinter, T. A. W., & van Saase, J. L. C. M. (2010). Conscious thought beats deliberation without attention in diagnostic decision-making: At least when you are an expert. *Psychological Research*, *74*(6), 586–592. doi: 10.1007/s00426-010-0281-8

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, *52*(6), 352–361. doi: 10.1016/j.jmp.2008.04.003

McCammon, I., & Hägeli, P. (2007). An evaluation of rule-based decision tools for travel in avalanche terrain. *Cold Regions Science and Technology*, *47*(1–2), 193–206. doi: 10.1016/j.coldregions.2006.08.007

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education*, *44*(1), 94–100. doi: 10.1111/j.1365-2923.2009.03507.x

Norman, G. R., Sherbino, J., Dore, K., Wood, T., Young, M., Gaissmaier, W., ... Monteiro, S. (2014). The etiology of diagnostic errors: A controlled trial of system 1 versus system 2 reasoning. *Academic Medicine*, *89*(2), 277–284. doi: 10.1097/ACM.0000000000000105

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240. doi: 10.1016/j.cogpsych.2012.03.003

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, United Kingdom: Cambridge University Press.

Pleskac, T. J., Cesario, J., & Johnson, D. J. (2018). How race affects evidence accumulation during the decision to shoot. *Psychonomic Bulletin & Review*, 1301–1330. doi: 10.3758/s13423-017-1369-6

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, *143*(5), 2000–2019. doi: 10.1037/xge0000013

Plummer, M., et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124).

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks.

Neural Computation, *20*(4), 873–922. doi: 10.1162/neco.2008.12-06-420

Rebitschek, F. G., Bocklisch, F., Scholz, A., Krems, J. F., & Jahn, G. (2015). Biased processing of ambiguous symptoms favors the initially leading hypothesis in sequential diagnostic reasoning. *Experimental Psychology*, *62*(5), 287–305. doi: 10.1027/1618-3169/a000298

Rebitschek, F. G., Krems, J. F., & Jahn, G. (2015). Memory activation of multiple hypotheses in sequential diagnostic reasoning. *Journal of Cognitive Psychology*, *27*(6), 780–796. doi: 10.1080/20445911.2015.1026825

Sherbino, J., Dore, K. L., Wood, T. J., Young, M. E., Gaissmaier, W., Kreuger, S., & Norman, G. R. (2012). The relationship between response time and diagnostic accuracy. *Academic Medicine*, *87*(6), 785–791. doi: 10.1097/ACM.0b013e318253acbd

Su, Y.-S., & Yajima, M. (2015). R2jags: Using r to run 'JAGS' [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=R2jags (R package version 0.5-7)

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology*, *28*(19), 3128–3135. doi: 10.1016/j.cub.2018.07.052

Thammasitboon, S., & Cutrer, W. B. (2013). Diagnostic decision-making and strategies to improve diagnosis. *Current Problems in Pediatric and Adolescent Health Care*, *43*(9), 232–241.

Todd, P. M., Gigerenzer, G., & ABC Research Group (Eds.). (2012). *Ecological rationality: Intelligence in the world*. New York, NY: Oxford University Press.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. doi: 10.1037/a0021765

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. doi: 10.1111/j.1467-9280.2008.02136.x

Zalaudek, I., Argenziano, G., Soyer, H. P., Corona, R., Sera, F., Blum, A., ... The Dermoscopy Working Group (2006). Three-point checklist of dermoscopy: An open internet study. *British Journal of Dermatology*, *154*(3), 431–437. doi: 10.1111/j.1365-2133.2005.06983.x

5 | The Ecological Rationality of the Wisdom of Crowds

Herzog, S. M., Litvinova, A., Yahosseini, K. S., Tump, A. N., & Kurvers, R. K. J. M. (2019).
In R. Hertwig, T. J. Pleskac, T. Pachur, & Center for Adaptive Rationality,
Taming uncertainty (pp. 245–262). Cambridge, MA: MIT Press.
https://doi.org/10.7551/mitpress/11114.003.0019

6 | Summary and Future Directions

Inconsistency in judgments or decisions within and between individuals is commonly seen as a reason for concern. Identical cases must be evaluated identically (among and within individuals; Kahneman, Rosenfield, Gandhi, & Blaser, 2016); consequently, when judgments are inconsistent, some of them must be wrong. Naturally, individuals have the urge to identify the most accurate judgment or best performing individual, neglecting that combining judgments can perform at least as well as the most accurate judgment or individual and sometimes even outperforms them (Soll & Larrick, 2009; Yaniv, 2004). This dissertation treated inconsistent judgments from a different perspective. Extending research on the wisdom of the inner crowd (Herzog & Hertwig, 2014), I showed in previously unexplored domains (i.e., confidence judgments of general knowledge and medical decisions) when conflicting judgments arise and how they can be used to one's benefit. Moreover, this work investigated cognitive dependency processes in sequential diagnostic reasoning tasks and explored a different procedure to reduce such dependencies. Finally, my colleagues and I reviewed four well-known aggregation strategies for outer crowds from an ecological perspective and pointed out the parallels and differences between aggregating judgments across and within individuals.

Chapter 2 addressed the accuracy of confidence judgments in two-alternative choice tasks. Despite the often bad reputation of subjective confidence judgments, confidence is frequently used in various real-world areas of decision making, including intelligence service (Betts, 1978; Mandel & Barnes, 2014; Mellers et al., 2014) and eyewitness (Wixted, Mickes, Dunn, Clark, & Wells, 2016) reports, the stock market, and medical diagnostics (Berner & Graber, 2008). How might confidence judgments be improved? Chapter 2 compared the performance of two strategies for aggregating an individual's conflicting confidence judgments in different statistical environments. Using analytical and simulation approaches and empirical data, the results showed that irrespective of the environment, averaging one's conflicting confidence judgments improves accuracy relative to first confidence judgments. Choosing the higher confidence judgment (i.e., maximizing) instead is risky; it harms accuracy for "wicked" items and only begins to outperform averaging once items are answered correctly 60% of the time or more. Put differently, items must be answered correctly above chance level (50% of the time) for maximizing to outperform averaging. This chapter provides at least three novel contributions. First, it offers the first qualitative and quantitative comparison of two competing strategies (i.e., averaging and maximizing) for aggregating confidence judgments. Second, it provides concrete, quantitative predictions for each strategy's performance. Third, it demonstrates the theoretical predictions in three empirical studies

drawn from different domains (general knowledge, demography, and perceptual decision making).

Chapter 3 addressed the questions of when experts change their mind and what to do about it. Inconsistency in expert judgment is often understood as a source of error (Kahneman et al., 2016) and can have serious consequences; for example, when a physician makes contradicting diagnoses in repeated decisions. To understand the conditions when experts change their mind, we investigated the relationship between inconsistency in decisions, confidence judgments, and the ambiguity of a case. Are experts more likely to change their mind when they are initially wrong? And can confidence judgments predict when experts change their mind? To answer these and more questions, my colleagues and I used a theoretical model based on the self-consistency model (Koriat, 2012a) and investigated the resulting predictions in two real-world high-stakes expert datasets. Our findings are threefold. First, the stronger the consensus among experts, the less likely they were to make a different decision the second time around, even if the consensus was wrong. This is further corroborated by the fact that cases with high expert agreement were diagnosed with high confidence, implying that when experts highly agree on a diagnosis that is wrong, they will confidently stick to the wrong diagnosis. Second, when fewer experts agreed with each other, their confidence judgments became less confident. Third, the less confident an expert's judgment, the more likely she was to change her mind when judging the same case again. Finally, whenever experts change their mind, our results suggest they should choose the decision with the higher confidence. This might seem to contradict the previous chapter's recommendation, but it is in line with the ecological rationality of the strategy's performance: Because experts are usually more likely to be correct than wrong, their confidence can be used as a signal for the accuracy of their decisions and hence as a guide for dealing with inconsistent decisions.

Chapter 4 moved from repeated judgments to the domain of sequential diagnostic decision making and investigated how the inner crowd can be harnessed to reduce cognitive dependency processes, such as a biased or less attentive mind. A majority of diagnostic errors can be attributed to a doctor's cognitive processes (Graber, Franklin, & Gordon, 2005; Hussain & Oestreicher, 2017; Norman & Eva, 2010; but see Sherbino et al., 2012). In sequential diagnostic decision making, for example, the order of the encountered evidence influences how consecutive pieces of evidence are evaluated (Kostopoulou, Mousoulis, & Delaney, 2009; Lange, Thomas, & Davelaar, 2012; Rebitschek, Bocklisch, Scholz, Krems, & Jahn, 2015; Rebitschek, Krems, & Jahn, 2015). Chapter 4 investigated whether sequential diagnostic procedures induce dependencies in cognitive processes between decisions in a sequence, taking into account the probabilistic structure of the environment. Additionally, my colleagues and I explored whether an alternative procedure that aimed to free one's mind from one's initial assessment by separating the assessments in time could reduce such dependencies. Our results show that cognitive dependencies primarily exist in an environment with correlated cues, but that the alternative procedure does not succeed in reducing such dependencies.

Finally, in Chapter 5 my colleagues and I took a broader perspective and reviewed four well-known aggregation strategies for outer crowds, that is, for groups of several individuals, and pointed out the parallels and differences between combining decisions in outer and inner crowds. We demonstrate how each aggregation strategy has its own ecological niche—there is no single strategy that outperforms every other strategy across

all statistical environments. The similarity between aggregation strategies for outer and inner crowds is, that the success of strategies depends on the similarity of the accuracy of the aggregated judgments, diversity of errors, and the ability to identify the better judgment. The difference is, that averaging judgments of different individuals usually results in higher gains because error diversity between different individuals' judgments is higher. In sum, aggregation strategies offer a powerful tool to reduce uncertainty. However, because the statistical properties of an environment are often unknown, individuals face a new type of uncertainty: Which strategy should be selected? In the absence of information about the statistical environment, we suggest two principles: (a) aggregate more rather than fewer judgments, and (b) use experience to adapt to the environment. Aggregating more judgments will balance against the risk of choosing the worst performing individual, while accumulated experience (e.g., through feedback) can help in becoming more selective in deciding whose judgments should be included.

In the next section I first discuss questions that were left open or newly emerged in this dissertation. Then I transition to guiding questions for future research, and conclude with connecting the insights from the wisdom of *human* crowds to a new type of crowd that is gaining increasing relevance, the crowd arising from human–machine interactions.

What Remains Open?

One natural question that arises from the work presented in Chapter 2 is whether the accuracy of decisions in two-alternative choice tasks can be improved through averaging or maximizing confidence judgments. In principle, whenever binary decisions differ, averaging and maximizing their associated confidence judgments will necessarily result in the same decision, and therefore those two strategies cannot differ in terms of the proportion of correct decisions.[1] To illustrate this with an example, assume an individual's first answer to the question "Sofia is the capital of: (a) Romania or (b) Bulgaria?" was "Romania" with 60% confidence. At the second assessment the individual changed her mind and answered "Bulgaria" with 100% confidence. Maximizing would result in choosing Bulgaria with 100% confidence. Likewise, assuming that answering Romania with 60% confidence is equivalent to answering Bulgaria with 40% confidence, averaging would result in choosing Bulgaria with 70% confidence. Koriat (2012b) demonstrated that selecting the decision with the higher confidence can improve the accuracy (from 81% to 82% of correct decisions) of an individual's decisions in "kind" environments. Our results show an effect on the accuracy of confidence judgments, yet no consistent effects on the accuracy of decisions when averaging or maximizing.[2] However, one necessary condition to improve the accuracy of decisions is that decisions must differ in the first place. Our participants, on average, changed their decisions in 12% to 22% of questions—putting an upper limit on any improvements one could see

[1] To see why, consider two confidence judgments (coded as the confidence in the correct decision). First consider the case of both being larger or both smaller than 0.5 (i.e., they imply the same decision): Both the more extreme confidence judgment as well as their average remain on that same side of 0.5 and therefore imply the same decision. If the two confidence judgments are on opposite sides of 0.5 (i.e., they imply different decisions), then the arithmetic average will be on the side of 0.5 with the more extreme confidence judgment; and, by definition, maximizing will also choose that very same decision.

[2] Study 3 constitutes the only exception, with second and averaged dialectical estimates showing a mean increase of 2 percentage points.

in terms of proportion of correct decisions. The reason we focused on the quality of the confidence judgments and less on the accuracy of the decisions themselves lies in the fact that confidence itself is an important ingredient for decision making in that it will, among other things, determine whether people will act on decisions (e.g., bets) or advice and whether they will consult additional information or advisors. That is, not only is confidence an additional aspect of decisions, it also has tangible effects on decision making.

Given the tight connection between Chapter 2 and 3, one might wonder why we did not directly compare averaging and maximizing confidence judgments in expert judgments. Unfortunately, experts confidence judgments were either provided on a two-point or five-point verbal scale. To measure the performance of averaging and maximizing in terms of the Brier score, we would have needed confidence judgments on a six-point scale to translate them to a half-range [50%, 60%, 70%,... ,100%] probability scale.

Chapter 4 introduced new questions regarding the ecological rationality of dependencies in cognitive processes during sequential diagnostic decision making. Researchers have found that the majority of medical errors can be attributed to cognitive processes (Graber et al., 2005), and that in sequential diagnostic procedures, initial assessments influence later assessments (Kostopoulou et al., 2009; Lange et al., 2012; Rebitschek, Bocklisch, et al., 2015; Rebitschek, Krems, & Jahn, 2015). This chapter investigated whether the sequential nature of diagnostic procedures induces such dependencies between assessments, taking into account the probabilistic structure of the environment, and explored methods to reduce cognitive dependencies. Results show that cognitive dependency processes appeared primarily in an environment where cues were correlated, suggesting that the probabilistic structure of the environment moderates these processes. Our findings raise the question of to what extent it is adaptive to integrate the underlying statistical structure of the environment into one's decision making process. Future research should, therefore, answer this normative question, taking into account not only the accuracy of successive subdecisions in a sequence but also the final diagnosis. Depending on the answers to these questions, further steps can be taken to reduce diagnostic errors.

Guiding Questions for Future Research

Guiding questions for future studies can be broadly divided into two streams, a cognitive stream that gets to the bottom of the phenomenon of the inner crowd, and another, more applied stream that explores possible implementations of inner crowd strategies in various areas of real-world decision making. Concerning the cognitive understanding, to date the inner crowd has been demonstrated on a behavioral level only. Advancements in neuroimaging techniques, such as magnetic resonance imaging (MRI), provides additional tools to study the inner crowd on a neurological level (De Martino et al., 2018). Recent technical and computational innovations in ultra high field MRI scanners (with 7 Tesla and above) made it possible for imaging tools to operate in vivo at the mesoscopic scale and fostered the use of functional MRI for computational modeling of neural networks. Such advancements make it possible to study neural computations at a submillimeter scale, implying a spatial resolution of cortical layers, cortical columns, and cortical nuclei—which previously could only be achieved with invasive techniques (De Martino et al., 2018; Haupt et al., 2017). These developments

open new doors to study the wisdom of the inner crowd. Neurologically, distinct information should be represented in different cortical regions, or distinct cortical connections. For example, studies using neural data have been able to reconstruct the visual field during perception and mental imagery of four different letter shapes (Senden, Emmerling, Van Hoof, Frost, & Goebel, 2018) and identify a specific story a participant was reading (Dehghani et al., 2017). If it is possible to decode a specific letter that an individual is imagining, or a specific story to which an individual is listening, would it also be possible to identify whether an individual is making use of an inner crowd when making judgments? How might imaging studies differentiate between inconsistent judgments resulting from different sources of information—the key driving force behind the wisdom of the inner crowd—and those coming from a stochastic process after information has been retrieved? To answer these questions we need not only high spatial resolution but also high temporal resolution in imaging techniques. Let me speculate about possible implications for research on the wisdom of the inner crowd. The application of newly emerging imaging techniques in combination with cognitive models could further advance the understanding of the wisdom of the inner crowd and provide neurological proof of concept: Future research could investigate whether inconsistent judgments are the result of diverse samples of information (e.g., represented by distinct neural networks at the time of sampling information) or of a perhaps stochastic generation of judgments after samples have been retrieved (e.g., represented by distinct neural networks after sampling). This research could then be used to understand (a) the underlying process of mind reversals, (b) in what situations individuals are more likely to change their mind, and (c) to design tools that facilitate retrieval of diverse information.

Applied research should identify further areas of real-world decision making, beyond the medical domain, where inner crowd strategies can be implemented. There are at least three situations that lend themselves to the application of inner crowd strategies: (a) when it is difficult to collect opinions of several individuals, for example, because of limited budget to consult more experts, time constraints, or logistic limitations; (b) when it is unlikely to recognize previously evaluated cases as such, for example, in radiology or dermoscopy, where images can look fairly similar throughout a set of cases; and (c) when dealing with sensitive cases, where, for example, security concerns prevent the sharing of information across multiple individuals.

The Human–Machine Crowd

The focus of this dissertation was, broadly speaking, on the wisdom of *human* crowds. However, with the progress in artificial intelligence (AI) a fairly new type of crowd is gaining increasing relevance: the crowd arising from human–machine interactions. How could the insights from the wisdom of human crowds be transferred to human–machine crowds? In the following sections I briefly review what AI is contributing to human decision making in a specific area: the health care sector. Then, I discuss possible alternatives of how human–machine judgments could be combined and conclude with a final thought about what the human mind should contribute to AI systems.

In the health care sector, AI is assisting clinical decision making in various areas (e.g., radiology, oncology,

pathology, and brain imaging) and applications (e.g., drug discovery, patient monitoring, medical diagnostics, imaging, and hospital management; Hosny, Parmar, Quackenbush, Schwartz, & Aerts, 2018). Within radiology, radionomic studies (i.e., studies of radiographic images coupled with clinical outcomes; Lambin et al., 2012) have employed deep learning algorithms that automatically learn feature representations from example images (Litjens et al., 2017) and thus help in the interpretation of the phenotypic characteristics of human tissues (Shen, Wu, & Suk, 2017). These technological advances offer a large potential to improve disease diagnoses and to reduce medical errors. In some cases radiologists have on average only 3–4 seconds to interpret an image to meet workload demands (McDonald et al., 2015). Additionally, radiologists not only operate under time pressure, but also have to make decisions under incomplete evidence and perceptual uncertainty (Fitzgerald, 2001), making diagnostic errors almost inevitable. Further investing in and incorporating AI as a tool to assist in clinical decision making can thus increase efficiency and reduce errors.

From the wisdom of crowds perspective at least one important question arises: How should human and machine judgments be combined? The answer to this question will differ depending on the field of decision making and the statistical environment. In the health care sector future research should take at least three alternatives into account. First, from a purely objective point of view, one could argue that, in order to reduce medical errors, future studies should investigate which statistical aggregation strategy (e.g., confidence rule, best-member rule, averaging) for combining medical experts' and AI judgments performs well in which environment. However, purely statistical aggregation approaches will likely face legal constraints, because many AI models cannot be held accountable. Despite the growing success of AI, one major downside of relying on currently popular AI models, such as deep learning algorithms, is that they operate as a "black box," meaning that we have little to no insight into how the algorithm extracts information and learns to make categorizations. Such algorithms are called "deep" because they consist of multiple layers of information processing. Often the layers between the input and the output layer are described as "hidden layers" because their information processing is opaque. The uncertainty about how exactly such models draw their conclusions makes it difficult to predict errors or to correct bugs in the software. The inherent complexity and lack of transparency of such black box algorithms is what makes investigating and verifying their mathematical reasoning a major ongoing challenge (Ford & Price, 2016; Pasquale, 2015). A second alternative could be to use AI models as an augmented inner crowd, potentially providing information that might have stayed undiscovered by the human eye, but the expert remains autonomous (Hosny et al., 2018). Accordingly, experts would control, supervise, maintain, and optimize AI models—a concept called *human-in-the-loop* (HITL; Allen, Guinn, & Horvtz, 1999; Rahwan, 2018; Sheridan, 2006). Third, a perhaps less obvious approach could be to integrate patients' preferences into the final decision making process. Some individuals might dislike the idea of a machine influencing their diagnosis, while others have more trust in machines than in medical experts. Therefore, another alternative could be to give the patient a voice in deciding between the expert, AI, or a combination thereof.

I want to conclude by touching upon a final thought. In the above section I pointed out how AI can augment the human mind. In the final section, I want to bring up the question of what parts of the human mind should to be planted into AI algorithms. Going beyond the health care sector, newly emerging AI systems, such

as unmanned combat drones (White, 2003), autonomous vehicles (Endsley, 2018; Pendleton et al., 2017), and news-filtering or credit-scoring algorithms have broad implications on a societal level (Helbing, in press; Rahwan, 2018)—they can influence political beliefs, economic development and even the life or death of large groups of people. Should collateral damage be tolerated in autonomous warfare? Or should a self-driving car in an inevitable tradeoff scenario save the lives of passengers or pedestrians? These questions cannot be answered from a purely technical engineering point of view, but have to take human ethical preferences into account. Many scholars and policy makers argue that AI systems, influencing the political beliefs or even lives of groups of individuals, need to embed the values of the society in which they operate (Rahwan, 2018). The challenge, however, is that often the society does not know their values or cannot easily formulate ethical principles to guide machine behavior. One attempt to close this gap was undertaken by a large-scale societal study that simply asked citizens from 233 countries which outcome they would prefer, such as saving the lives of the elderly or the young in an inevitable self-driving car accident (Awad et al., 2018)? This study offers interesting insights into human ethical preferences regarding machine behavior. But even if the society would agree, that for example rather the lives of the young than the elderly should be saved, would it make it right? Should we plant this ethical preference into machine behavior? And do we need deterministic answers to these type of questions?

References

Allen, J., Guinn, C. I., & Horvtz, E. (1999). Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, *14*(5), 14–23. doi: 10.1109/5254.796083

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59. doi: 10.1038/s41586-018-0637-6

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5 Suppl), S2–S23. doi: 10.1016/j.amjmed.2008.01.001

Betts, R. K. (1978). Analysis, war, and decision: Why intelligence failures are inevitable. *World Politics*, *31*, 61–89. doi: 10.2307/2009967

Dehghani, M., Boghrati, R., Man, K., Hoover, J., Gimbel, S. I., Vaswani, A., ... others (2017). Decoding the neural representation of story meanings across languages. *Human brain mapping*, *38*(12), 6096–6106. doi: doi.org/10.1002/hbm.23814

De Martino, F., Yacoub, E., Kemper, V., Moerel, M., Uludag, K., De Weerd, P., ... Formisano, E. (2018). The impact of ultra-high field MRI on cognitive and computational neuroimaging. *NeuroImage*, *168*, 366–382. doi: 10.1016/j.neuroimage.2017.03.060

Endsley, M. R. (2018). Situation awareness in future autonomous vehicles: Beware of the unexpected. In *Congress of the International Ergonomics Association* (pp. 303–309).

Fitzgerald, R. (2001). Error in radiology. *Clinical Radiology*, *56*(12), 938–946. doi: 10.1053/crad.2001.0858

Ford, R. A., & Price, W. N. I. (2016). Privacy and accountability in black-box medicine. *Michigan Telecommunications and Technology Law Review*, *23*(1), 1493–1499.

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine*, *165*(13), 1493–1499. doi: 10.1001/archinte.165.13.1493

Haupt, D., Vanni, M. P., Bolanos, F., Mitelut, C., LeDue, J. M., & Murphy, T. H. (2017). Mesoscale brain explorer, a flexible python-based image analysis and visualization tool. *Neurophotonics*, *4*(3), 031210. doi: 10.1117/1.NPh.4.3.031210

Helbing, D. (in press). Societal, economic, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies. In D. Helbing (Ed.), *Towards digital enlightenment: Essays on the dark and light sides of the digital revolution* (pp. 47–72). Cham, Switzerland: Springer International.

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506. doi: 10.1016/j.tics.2014.06.009

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*(8), 500–510. doi: 10.1038/s41568-018-0016-5

Hussain, A., & Oestreicher, J. (2017). Clinical decision-making: Heuristics and cognitive biases for the ophthalmologist. *Survey of Ophthalmology*, *63*(1), 119–124. doi: 10.1016/j.survophthal.2017.08.007

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*, *94*(10), 38–46.

Koriat, A. (2012a). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. doi: 10.1037/a0025648

Koriat, A. (2012b). When are two heads better than one and why? *Science*, *336*(6079), 360–362. doi: 10.1126/science.1216549

Kostopoulou, O., Mousoulis, C., & Delaney, B. (2009). Information search and information distortion in the diagnosis of an ambiguous presentation. *Judgment and Decision Making*, *4*(5), 408–418.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., ... Dekker, A. (2012). Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, *48*(4), 441–446. doi: 10.1016/j.ejca.2011.11.036

Lange, N. D., Thomas, R. P., & Davelaar, E. J. (2012). Temporal dynamics of hypothesis generation: The influences of data serial order, data consistency, and elicitation timing. *Frontiers in Psychology*, *3*, 215–231. doi: 10.3389/fpsyg.2012.00215

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. doi: 10.1016/j.media.2017.07.005

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(30), 10984–10989. doi: 10.1073/pnas.1406138111

McDonald, R. J., Schwartz, K. M., Eckel, L. J., Diehn, F. E., Hunt, C. H., Bartholmai, B. J., ... Kallmes, D. F. (2015). The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic Radiology*, *22*(9), 1191–1198. doi: 10.1016/j.acra.2015.05.007

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. doi: 10.1177/0956797614524255

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education*, *44*(1), 94–100. doi: 10.1111/j.1365-2923.2009.03507.x

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Camebridge, MA: Harvard University Press.

Pendleton, S. D., Andersen, H., Du, X., Shen, X., Meghjani, M., Eng, Y. H., ... Ang, M. H. (2017). Perception, planning, control, and coordination for autonomous vehicles. *Machines*, *5*(1), 6. doi: 10.3390/machines5010006

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. doi: 10.1007/s10676-017-9430-8

Rebitschek, F. G., Bocklisch, F., Scholz, A., Krems, J. F., & Jahn, G. (2015). Biased processing of ambiguous symptoms favors the initially leading hypothesis in sequential diagnostic reasoning. *Experimental Psychology*, *62*(5), 287–305. doi: 10.1027/1618-3169/a000298

Rebitschek, F. G., Krems, J. F., & Jahn, G. (2015). Memory activation of multiple hypotheses in sequential diagnostic reasoning. *Journal of Cognitive Psychology*, *27*(6), 780–796. doi: 10.1080/20445911.2015.1026825

Senden, M., Emmerling, T. C., Van Hoof, R., Frost, M., & Goebel, R. (2018). Reconstructing and decoding imagined letters from early visual cortex using ultra-high field fMRI. *bioRxiv*, 277020. doi: 10.1101/277020

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Sherbino, J., Dore, K. L., Wood, T. J., Young, M. E., Gaissmaier, W., Kreuger, S., & Norman, G. R. (2012). The relationship between response time and diagnostic accuracy. *Academic Medicine*, *87*(6), 785–791. doi: 10.1097/acm.0b013e318253acbd

Sheridan, T. (2006). Supervisory control. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (Vol. 3rd ed, pp. 1025–1052). Hoboken, NJ: Wiley.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805. doi: 10.1037/a0015145

White, A. (2003). The human–machine partnership in UCAV operations. *The Aeronautical Journal*, *107*(1069), 111–116. doi: 10.1017/S0001924000013786

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(2), 304–309. doi: 10.1073/pnas.1516814112

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. doi: 10.1111/j.0963-7214.2004.00278.x

Appendices

A | Supplementary Material to Chapter 2: "How the "Wisdom of the Inner Crowd" Can Boost Accuracy of Confidence Judgments"

A1 Conditions Under Which Averaging Has a Smaller Expected Brier Score Than Maximizing

We consider a two-alternative forced-choice paradigm where a decision maker decides twice about the same item, that is, renders a first and a second decision concerning the same question. The two decisions either coincide or not. Furthermore, for each of the two decisions the decision maker also provides a confidence judgment (half-range, that is, the subjective probability of having made the correct decision, ranging between .5 and 1). We want to specify the conditions for which *averaging* (i.e., simply aggregating the two confidence judgments using the arithmetic mean) has a smaller expected Brier score (Brier, 1950) than *maximizing* (i.e., choosing the option with the higher associated confidence and reporting that confidence).

To investigate this analytically, we use a very general model that postulates for a particular item (1) the probability P that the *high*-confidence choice is correct, (2) the confidence C_H in this *high*-confidence choice, (3) the confidence C_L in the other, *low*-confidence choice, and (4) whether the high- and low-confidence choices are the same. The model makes no cognitive assumptions but merely restricts the admissible range of the three variables by making the following two assumptions. First, $0 < P < 1$. Second, $.5 < C_L < C_H < 1$; that is, the high-confidence judgment needs to be strictly larger than the low-confidence judgment and they are both expressed on a half-range probability scale.

The Brier score is the mean squared error across probability forecasts, and for a single item it can be expressed as $B = (o - f)^2$, where f is the probability forecast that an event o happens. Event o either happens ($o = 1$) or it does not ($o = 0$). In a two-alternative forced-choice paradigm the event o can be interpreted as whether the decision is correct ($o = 1$) or incorrect ($o = 0$). To derive the expected Brier scores for averaging and maximizing in this model, it is convenient to distinguish between the case when the two decisions differ and the case when the two decisions are the same and then to develop the equations separately for those two cases. Note that we define the event o as whether the high-confidence choice is correct.

Case 1: The Two Decisions Differ

Maximizing's expected Brier score $B_M^{different}$ is

$$E(B_M^{different}) = P(1 - C_H)^2 + (1 - P)(0 - C_H)^2 ,$$

whereas averaging's expected Brier score $B_A^{different}$ is

$$E(B_A^{different}) = P(1 - \frac{C_H + (1 - C_L)}{2})^2 + (1 - P)(0 - \frac{C_h + (1 - C_L}{2})^2$$

Because in Case 1 the low-confidence choice is the opposite of the high-confidence choice, we re-express the confidence in the low-confidence choice in terms of the confidence that the high-confidence choice is correct (i.e., we need to use $1 - C_L$ for the low-confidence choice).

Now we solve the following system of three inequalities (i.e., averaging having a lower Brier score than maximizing plus the two assumptions of the model):

$$\begin{cases} E(B_A^{different}) < E(B_M^{different} \\ 0 < P < 1 \\ 0.5 < C_L < C_H < 1 \end{cases} ,$$

which results in the following four conditions satisfying the above system of inequalities:

$$\begin{cases} 0 < P \leq \frac{1}{2} \text{ and } \frac{1}{2} < C_H < 1 \text{ and } \frac{1}{2} < C_L < C_H \\ \frac{1}{2} < P \leq \frac{3}{4} \text{ and } \frac{1}{6}(8P - 1) < C_H < \frac{1}{2}(4P - 1) \text{ and } \frac{1}{2} < C_L < -4P + 3C_H + 1 \\ \frac{1}{2} < P \leq \frac{3}{4} \text{ and } \frac{1}{2}(4P - 1) \leq C_H < 1 \text{ and } \frac{1}{2} < C_L < C_H \\ \frac{3}{4} < P < \frac{7}{8} \text{ and } \frac{1}{6}(8P - 1) < C_H < 1 \text{ and } \frac{1}{2} < C_L < -4P + 3C_H + 1 \end{cases}$$

At least two insights can be gained from those four solutions. First, Solution 1 shows that an item being wicked is sufficient for averaging to always outperform maximizing; note that the second and third parts of Solution 1 merely restate the model's assumptions about the confidence judgments. Second, Solutions 2, 3, and 4 show the conditions under which averaging outperforms maximizing when $.5 < P < \frac{7}{8}$ (i.e., a very difficult to moderately difficult kind item). Yet these conditions are complicated and depend on the particular relationships between P, C_H, and C_L. However, since none of the four solutions represent items for which $P \geq \frac{7}{8}$, this implies that for such very easy kind items ($P \geq \frac{7}{8}$), averaging will always have a worse Brier score than maximizing.

Case 2: The Two Decisions Are the Same

Because maximizing's confidence depends on only C_H (and not on C_L), maximizing's expected Brier score B_M^{same} is the same as $B_M^{different}$ in Case 1:

$$E(B_M^{same}) = P(1 - C_H)^2 + (1 - P)(0 - C_H)^2 ,$$

whereas averaging's expected Brier score B_A^{same} is now

$$E(B_A^{same}) = P(1 - \frac{C_H + C_l}{2})^2 + (1-p)(0 - \frac{C_H + C_L}{2})^2 .$$

Now we solve the following system of three inequalities (i.e., averaging having a lower Brier score than maximizing plus the two assumptions of the model):

$$\begin{cases} E(B_A^{same}) < E(B_M^{same}) \\ 0 < P < 1 \\ 0.5 < C_L < C_H < 1 \end{cases},$$

which results in the following four conditions satisfying the above system of inequalities:

$$\begin{cases} 0 < P \leq \frac{1}{2} \text{ and } \frac{1}{2} < C_H < 1 \text{ and } \frac{1}{2} < C_L < C_H \\ \frac{1}{2} < P \leq \frac{7}{8} \text{ and } P < C_H < \frac{1}{6}(8P-1) \text{ and } 4P - 3C_H < C_L < C_H \\ \frac{1}{2} < P \leq \frac{7}{8} \text{ and } \frac{1}{6}(8P-1) \leq C_H < 1 \text{ and } \frac{1}{2} < C_L < C_H \\ \frac{7}{8} < P < 1 \text{ and } P < C_H < 1 \text{ and } 4P - 3C_H < C_L < C_H \end{cases}$$

At least two insights can be gained from those four solutions. First, Solution 1 shows that an item being wicked is, again, sufficient for averaging to always outperform maximizing; note that, again, the second and third parts of Solution 1 merely restate the model's assumptions about the confidence judgments. Second, Solutions 2, 3, and 4 show that for kind items of any difficulty level ($.5 < P < 1$) there are always conditions for which averaging can outperform maximizing. Or, phrased differently, for kind items there are no sufficient conditions for which maximizing always outperforms averaging that depend only on P (unlike in Case 1 discussed above where $P \geq \frac{7}{8}$ is a sufficient condition). These conditions, however, again are complicated and depend on the particular relationships between P, C_H, and C_L.

Summarizing Across Cases 1 and 2

First, for a wicked item (i.e., $P < .5$), averaging always has a better expected Brier score than maximizing—irrespective of whether the low-confidence choice is also incorrect or instead correct. Second, for a kind item (i.e., $P > .5$) the conditions are more complicated and depend on whether the low-confidence choice is also correct or instead incorrect. When the high-confidence choice is very likely to be correct (i.e., $P \geq \frac{7}{8}$, that is, a very easy kind item) but the low-confidence choice is incorrect, maximizing always has a better expected Brier score than averaging. In contrast, when both the low- and high-confidence choices are correct, there are no sufficient conditions for which maximizing always has a better expected Brier score than averaging that depend only on P. There are a series of conditions that specify for particular relationships between P, C_H, and C_L whether averaging or maximizing will have a better expected Brier score.

A2 Items Used in Study 3

Table A1. Items used in main study.

Question	Answer (a)	Answer (b)
When was the zipper invented?	Before 1920*	After 1920
Which country send the first terrestrial satellites to the orbit?	The Soviet Union*	USA
The first air mail was set up in:	England*	Germany
Kurt Gödel was:	A composer	A mathematician*
The number of leukocytes in the healthy human blood is:	Less than 4000/mm3	More than 4000/mm3*
Mao Zedong was born Before	1900*	After 1900
When was discovered the magnetic North Pole?	1866	1831*
Which of these fruits contains fat?	The lemon*	The bell pepper
Edgar Allan Poe was:	American*	Englishman
What does the word "hecatomb" mean?	Sacrifice to the idols*	Early Christian sepulchre/tomb
Who was born first?	Immanuel Kant*	Wolfgang Amadeus Mozart
Where can we find "fibrin"?	In a cell nucleus	In blood*
Who wrote the play "Liebelei"?	Arthur Schnitzler*	Franz Grillparzer
What's the name of the Bolivian capital?	La Paz*	Bogota
Where do the Betschuans live?	In Africa*	In Asia
Manuel da Falla was a:	Composer	Race driver
Sofia is the Capital of:	Romania	Bulgaria*
Who was the tutor of Alexander the Great?	Aristotle*	Plato
A meridian is a:	Circle of latitude	Circle of longitude*
Which metal melts down at a lower temperature?	Zinc	Tin*
Saskatchewan is (was) a state of:	The Soviet Union	Canada*
Weisherbst (Roséwine) is extracted from:	Red grapes*	White grapes
How long is the gestation time of an elephant?	22 months*	18 months
The first coffeehouse in Vienna was founded in:	1685*	1679
How many % from the whole Swiss grain production to the cattle eat?	More than 50%*	Less than 50%

Note. Correct answers are indicated with an asterisk.

A3 Decomposition of Overall Accuracy in the Simulation

Homogeneous Environments

Apart from overall accuracy (i.e., Brier score; see main text), confidence judgments can be evaluated along several dimensions of accuracy. To investigate how calibration and resolution contribute to overall accuracy (in terms of the Brier score) and how they are influenced by the environment and the dependency among knowledge sources, we decomposed the Brier score using the covariance decomposition (Yates, 1990). The three main components of the covariance decomposition are bias, slope, and scatter. The formula for the covariance decomposition of the Brier score is: $\overline{PS} = VI + \text{bias}^2 + VI(\text{slope})(\text{slope} - 2) + \text{scatter}$ (Yates, 1990),

where VI is the variability index, which is the variance of the event's probability, here the sample variance of the proportion of correct decisions. The bias score measures the difference between the average confidence judgment and the proportion of correct judgments:

$$bias = \overline{conf} - P(C)\,.$$

A positive bias score indicates overconfidence, or in other words, that individuals overestimate their probability of being correct. As one would expect by the definition of maximizing, the simulation study illustrates (Figure A1) that maximizing increases *bias* irrespective of the environment, whereas averaging has no effect on the *bias* score. As the dependency in knowledge sources (i.e., correlation r) increases, the effects of maximizing and averaging decrease. Because this pattern is stable across all reported measures (see Figure A7), we do not refer further to the effect of r.

The slope score is a measure of resolution. More specifically, it quantifies the difference between the average confidence in correct decisions and average confidence in wrong decisions:

$$slope = \overline{conf}_{correct} - \overline{conf}_{wrong}\,.$$

A positive slope score indicates the ability of confidence judgments to discriminate between correct and wrong decisions. As expected by the design of the simulation (Figure A1), confidence in the correct decision increases ($slope > 0$) and confidence in the wrong decision decreases as $p(C)$ increases in kind environments [i.e., $p(C) > 0.5$], implying that confidence judgments increasingly discriminate better between correct and wrong decisions (i.e., *positive resolution*). In wicked environments [i.e., $p(C) < 0.5$], however, the slope score falls below 0, implying a worse-than-chance ability of confidence judgments to discriminate between correct and wrong decisions (i.e., *negative resolution*). That is, as items become more wicked, confidence in the wrong decision increases and confidence in the correct decision decreases as $p(C)$ decreases. This pattern of results is consistent with Koriat's (2012a) consensuality principle, which states that because confidence is based on an assessment of how clearly a set of available cues agree with the selected response, confidence will be correlated with the strength of the majority belief (*consensuality*). Our simulation analysis thus complements Koriat's analysis by illustrating the consensuality principle in a quantitative manner.

The scatter score is the weighted average of the variances of confidence judgments for correct and wrong decisions:

$$scatter = \frac{n_{correct} var(conf_{correct} + n_{wrong} var(conf_{wrong}}{n_{correct} + n_{wrong}}$$

and represents the variability (i.e., random error) of confidence judgments, whereby larger scores indicate greater random error. Random error in confidence judgments was highest in environments where the probability of answering an item correctly is .5 (Figure A1). Furthermore, in line with Error theory (Wallsten & Diederich, 2001), both averaging and maximizing confidence judgments reduce scatter, and we observed a stronger effect for averaging compared to maximizing.

We additionally calculated a standardized measure of discrimination based on slope and scatter:

$$DI' = \frac{\text{slope}}{\sqrt{\text{scatter}}}$$

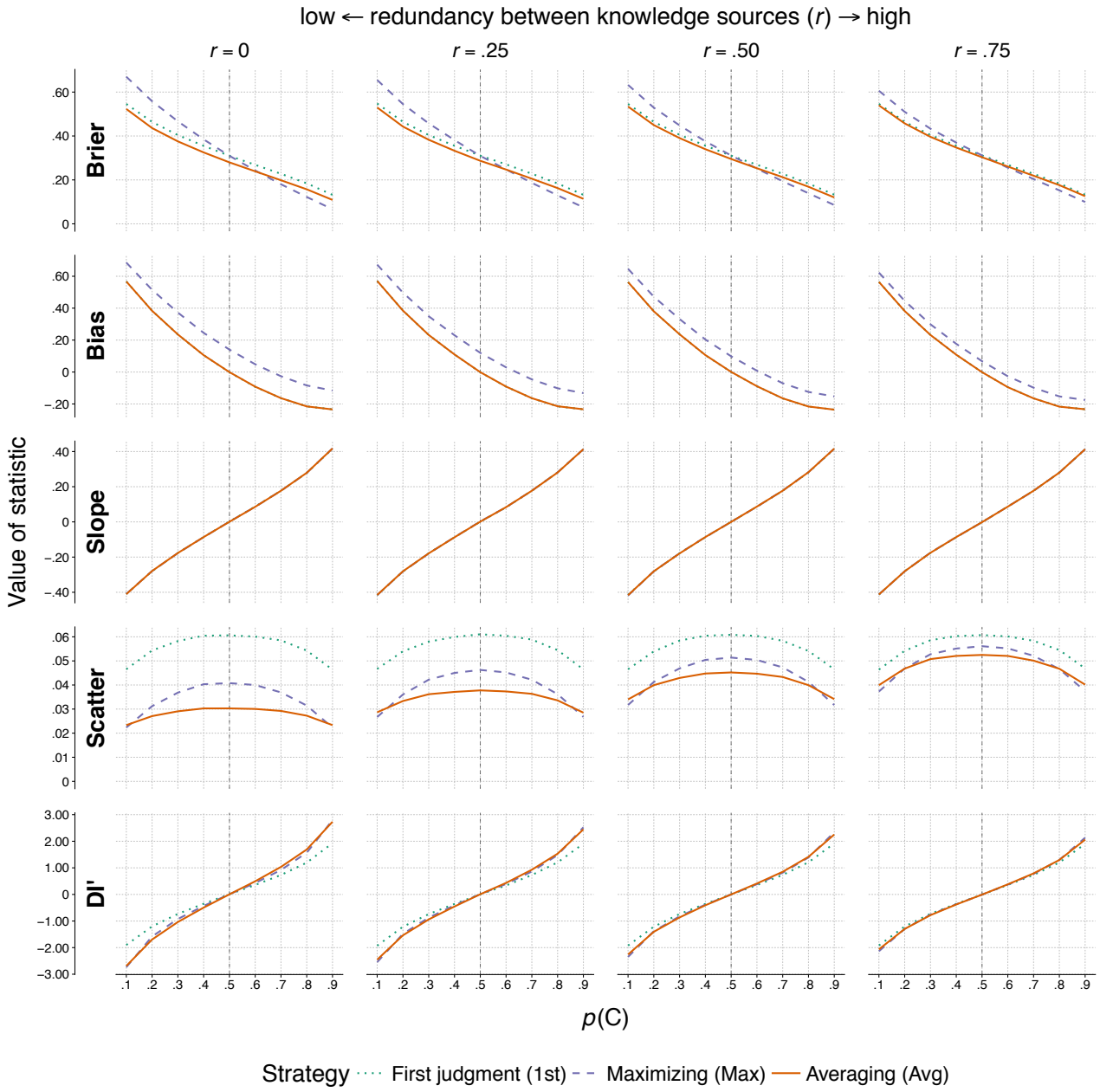


Figure A1. Decomposition of the Brier score of simulated strategies into the score components (rows). Columns (from left to right) correspond to increasingly more redundant knowledge sources underlying both confidence judgments (correlation values r). The x axes show the probability of being correct, where values of $p(C) > .5$ represent increasingly kinder environments and values of $p(C) < .5$ represent increasingly more wicked environments. The y axis of each row depicts the value of the corresponding statistic. In the bias panel, lines corresponding to first judgments are overplotted by averaging. In the slope panel, all strategies have the same values and are thus overplotting each other. DI' is a standardized measure of resolution, combining slope and scatter.

The DI' score can be interpreted along the same lines as the slope score, with $DI' > 0$ indicating positive discrimination and $DI' < 0$ meaning negative discrimination. Comparing first judgments to averaged and maximum judgments, averaging and maximizing amplify discrimination by increasing positive resolution for kind item types, that is, $p(C) > .5$ and by increasing negative resolution for wicked item types, that is, $p(C) < .5$ (Figure A1). Taken together, the results indicate that averaging outperforms maximizing on two out of the three components of the Brier score decomposition (bias and scatter), explaining why averaging outperforms maximizing overall in terms of the Brier score.

Heterogeneous Environments

In addition to the homogeneous environments (see main text), we also investigated various mixed environments, where the probability $p(C)$ of answering correctly differed across items (modelled as beta distributions; see Figure A2, panel A). We constructed the environments by orthogonally varying the mean μ of the *beta* distribution (values: [.1, .2, .3,... .9]), its variance σ^2 (values: [.01, .02, .03,... .08]), and the correlation r between the knowledge sources underlying the repeated confidence judgments from the same individual (values: [0, .25, .5, .75]). This resulted in 288 different environments. Other than assuming distributions of $p(C)$ (as compared to constant values for $p(C)$), the simulation procedure was identical to that described in the main text.

As in the homogeneous environments, also for heterogeneous environments we found that, as dependency in knowledge sources (i.e., correlation r) increased, the effects of averaging and maximizing on overall accuracy (Brier score) decreased (Figure A7). Therefore, in the following we only present results of the environments with zero dependency (i.e., $r = 0$).

Overall accuracy (Brier score)

In general, overall accuracy of confidence judgments improved (i.e., decreased Brier scores) with increasing mean μ (Figure A2, panel B). In wicked environments ($\mu < .5$) increasing variance σ^2 *reduced* the proportion of wicked items (Figure A2, panel C), but still harmed the Brier score of all strategies (increasing curves for all strategies in Figure A2, panel B). In contrast, in kind environments ($\mu > .5$) increasing variance σ^2 increased the proportion of wicked items (Figure A2, panel C) and improved the Brier score of first and averaged judgments (decreasing curves), but continued to harm the Brier score of maximizing (increasing curves).

Similar to the homogeneous environments, averaging improved overall accuracy (i.e. lower Brier scores; see Figure A2, panel B) relative to first judgments irrespective of the environment. Maximizing, in contrast, only improved overall accuracy in kind environments with $\mu \geq .6$ and harmed accuracy in environments with $\mu < .6$. Furthermore, in clearly kind environments with $\mu > .6$ maximizing outperformed averaging in low variance environments (e.g., $\mu = .7$, $\sigma^2 \leq .05$) but underperformed it in some of the high variance environments (e.g., $\mu = .7$, $\sigma^2 \geq .06$).

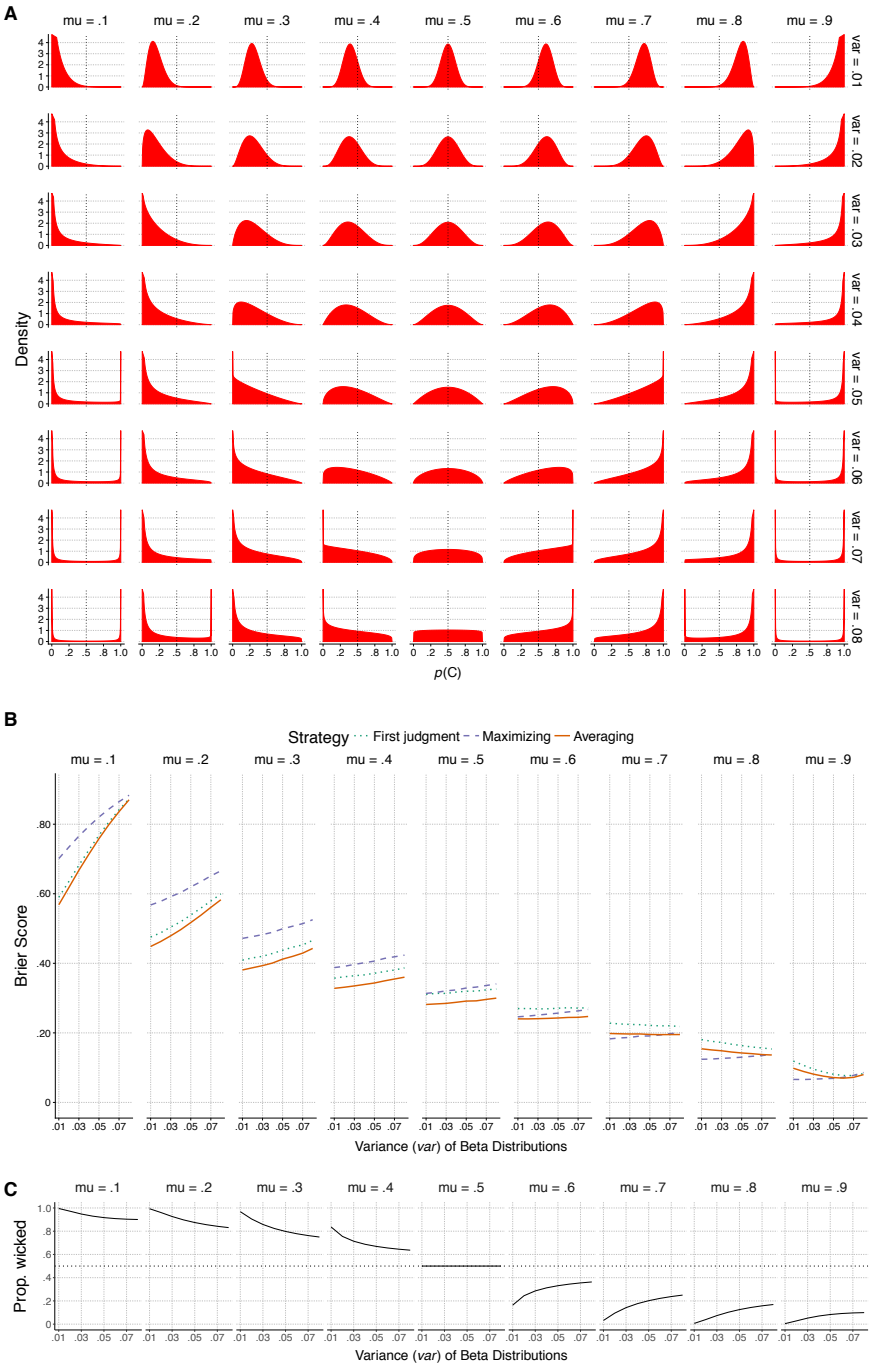


Figure A2. Brier scores of simulated strategies in heterogeneous environments. A. Heterogeneous environments varying according to the mean of $p(C)$ (μ: [.1, .2, .3,... .9]) and the variance of $p(C)$ (σ^2: [.01, .02, .03,... .08]). Columns correspond to increasingly more kind environments (μ), rows (from top to bottom) indicate increasing variance (σ^2). B. Brier scores (y axis) of simulated strategies in heterogeneous environments varying in μ (columns) and σ^2 (x-axes) for correlation $r = 0$. Averaging consistently improved Brier scores irrespective of the environment. Maximizing in contrast, harmed Brier scores for $\mu < .6$ and improves Brier scores only for $\mu \geq .6$. Maximizing outperforms averaging only for $\mu \geq .7$. Increasing variance harms all strategies for $\mu < .6$ (increasing curves), and benefits averaging and first judgments for $\mu > .6$. C. Proportion of wicked items in heterogeneous environments varying on μ (columns) and σ^2 (x axis). As variance increased, proportion of wicked items decreased in wicked environments ($\mu < .5$), and increased in kind environments ($\mu > .5$).

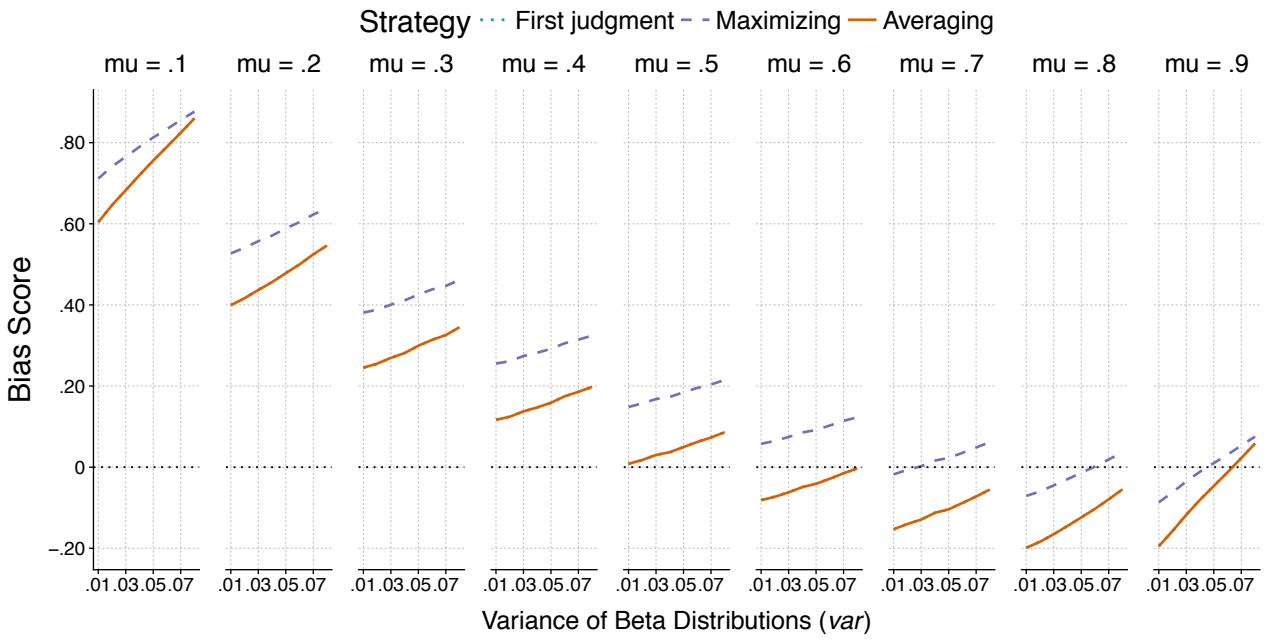


Figure A3. Bias scores (y axis) of simulated strategies in heterogeneous environments varying in the mean of $p(C)$ (μ; columns) and the variance of $p(C)$ (σ^2; x-axes) for correlation $r = 0$. Averaging is always overplotting first judgments.

Over- vs. underconfidence (bias score)

Bias scores decreased in increasingly kinder environments and increased with increasing variance for all strategies (Figure A3). Relative to first judgments, maximizing consistently increased bias, whereas averaging always has the same bias as first judgments. For averaging and first judgments, increasing variance σ^2 worsened (i.e., increased bias) for $\mu \leq .5$. For $\mu > .5$ averaging and first judgments were underconfident (i.e., negative bias score) and thus increasing variance σ^2 had a positive effect in that the increase in bias mostly resulted in the bias scores being closer to zero. Because maximizing had a higher bias score in general, this positive effect happens only for higher means and only up to moderate variances.

Over- vs. underconfidence (bias score)

Bias scores decreased in increasingly kinder environments and increased with increasing variance for all strategies (Figure A3). Relative to first judgments, maximizing consistently increased bias, whereas averaging always has the same bias as first judgments. For averaging and first judgments, increasing variance σ^2 worsened (i.e., increased bias) for $\mu \leq .5$. For $\mu > .5$ averaging and first judgments were underconfident (i.e., negative bias score) and thus increasing variance σ^2 had a positive effect in that the increase in bias mostly resulted in the bias scores being closer to zero. Because maximizing had a higher bias score in general, this positive effect happens only for higher means and only up to moderate variances.

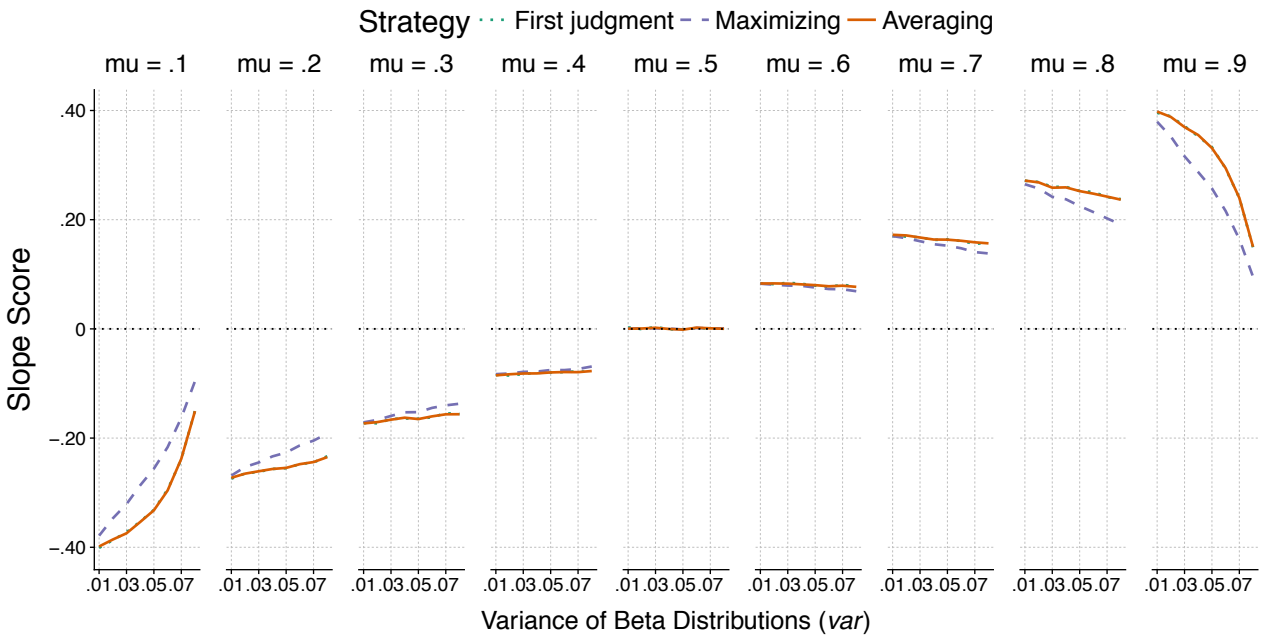


Figure A4. Slope scores (y axis) of simulated strategies in heterogeneous environments varying in the mean of $p(C)$ (μ; columns) and the variance of $p(C)$ (σ^2; x-axes) for correlation $r = 0$. Averaging is always overplotting first judgments.

Discrimination ability I (slope score)

In general, the kinder the environment (i.e., the higher μ), the better the slope (i.e., discrimination; Figure A4). In contrast, the effects of variance depended on the mean: In wicked environments ($\mu < .5$), where confidence discriminated between correct and incorrect decisions the wrong way around (i.e., reversed discrimination, that is, $slope < 0$), increasing variance reduced the size of the negative slope (i.e., the values of the slope increased and became thus less negative, that is, resulting smaller reversed discrimination). In contrast, in kind environments ($\mu > .5$), increasing variance reduced the size of the positive slopes. Relative to first judgments, averaging had no effect on the slope, while maximizing consistently differed, for better ($\mu > .5$) or worse ($\mu < .5$), from first and averaged judgments (except for $\mu = .5$).

Noise (scatter score)

Scatter scores were lowest in extremely kind (i.e., $\mu = .9$) or extremely wicked (i.e., $\mu = .1$) environments and highest in "ambiguous" environments (i.e., $\mu = .5$; Figure A5). Both averaging and maximizing improved scatter relative to first judgments—irrespective of the environment. An increasing variance had generally a negative effect on scatter except for extreme μ's, where the pattern was inversely U-shaped.

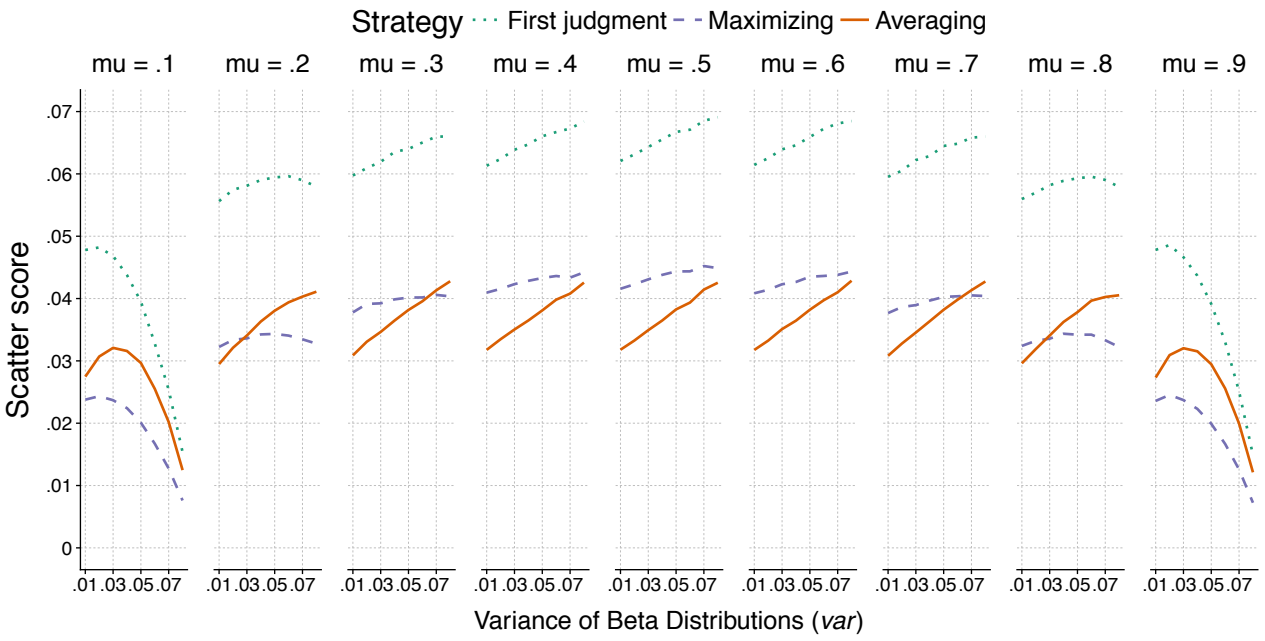


Figure A5. Scatter scores (y axis) of simulated strategies in heterogeneous environments varying in the mean of $p(C)$ (μ; columns) and the variance of $p(C)$ (σ^2; x-axes) for correlation $r = 0$. Averaging is always overplotting first judgments.

Discrimination ability II (DI')

DI' scores improved as environments became increasingly kinder (Figure A6). An increasing variance was beneficial in wicked environments (i.e., $\mu < .5$) and detrimental in kind environments (i.e., $\mu > .5$). Relative to first judgments, the DI' scores of both averaging and maximizing improved in kind environments, but worsend in wicked environments.

Dependency in knowledge sources

Figure A7 shows the Brier score for all environments when the dependency in knowledge sources increased (i.e., correlation values r: [0, .25, .5, .75]). As r increased, the differences between strategies decreased. Similar to the homogeneous environments, relative to first judgments, averaging always improved the Brier score, irrespective of the environment. Maximizing, in contrast, only improved the Brier score in kind environments ($\mu \geq .6$) and harmed the Brier score otherwise. Furthermore, in kind environments with μ [.7, .8, .9] maximizing outperformed averaging and first judgments. As variance increased (columns from left to right), differences between maximizing and averaging became smaller. When $\mu = .6$, averaging and maximizing performed similar in low variance environments (both improved performance on first judgments), however, as variance increased ($\sigma^2 \geq .03$), averaging started outperforming maximizing. When $\mu = .5$, averaging improved performance relative to first judgment, while maximizing harmed performance—especially when variance increased ($\sigma^2 \geq$

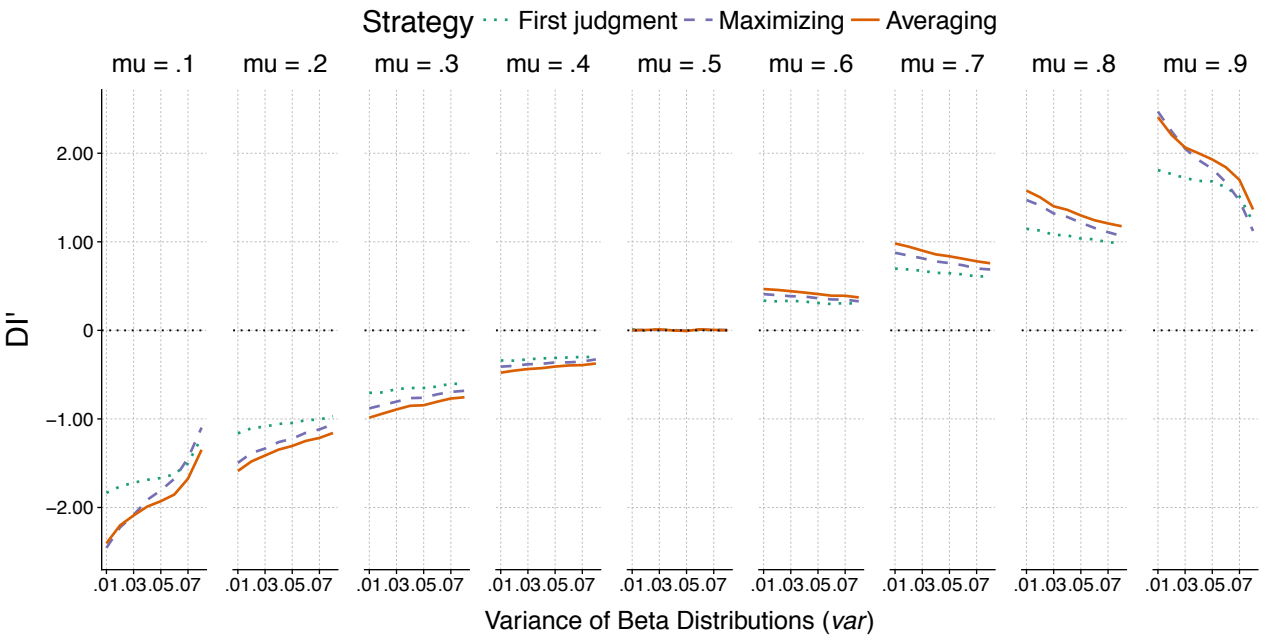


Figure A6. DI' scores (y axis) of simulated strategies in heterogeneous environments varying in the mean of $p(C)$ (μ; columns) and the variance of $p(C)$ (σ^2; x-axes) for correlation $r = 0$.

.03). This pattern continues for environments with $\mu < .5$. The effect of variance is most prominent in extreme environments with μ [.1, .2, .8, .9] where an increase in variance reduced the differences between all three strategies.

A4 Decomposition of Overall Accuracy in the Empirical Studies

To understand how the several dimensions of accuracy underlying the Brier score contribute to the changes in overall accuracy (see main text), we calculated, as in the simulation analysis, a Brier score decomposition (using the covariance decomposition Yates, 1990), which yields estimates of bias, slope, scatter, and DI' scores (refer to section A3 for more details on those scores).

The majority of effect sizes for the differences between averaging and maximizing with respect to (Table A2) and DI' (Table A3) are relatively small across studies (Figure A8)—with the two exceptions being an improved slope when averaging in the lines task ($d_{\text{lines}} = 0.47$, 95% highest density interval [0.16, 0.80]) and an impaired slope when maximizing in the shapes task ($d_{\text{shapes}} = -0.43$, 95% highest density interval [-0.80, 0.09]) of Study 2.

In contrast to the discrimination measures, the majority of scatter scores, a measure of random error, were greatly affected by both strategies. Averaging resulted in improved scatter scores (Table A9) in Studies 1 and 3, suggesting that improvements in overall accuracy were driven by a reduction of random error (except in Study 2). Similarly, maximizing reduced scatter scores in Studies 1 and 2, but these positive effects seem not to

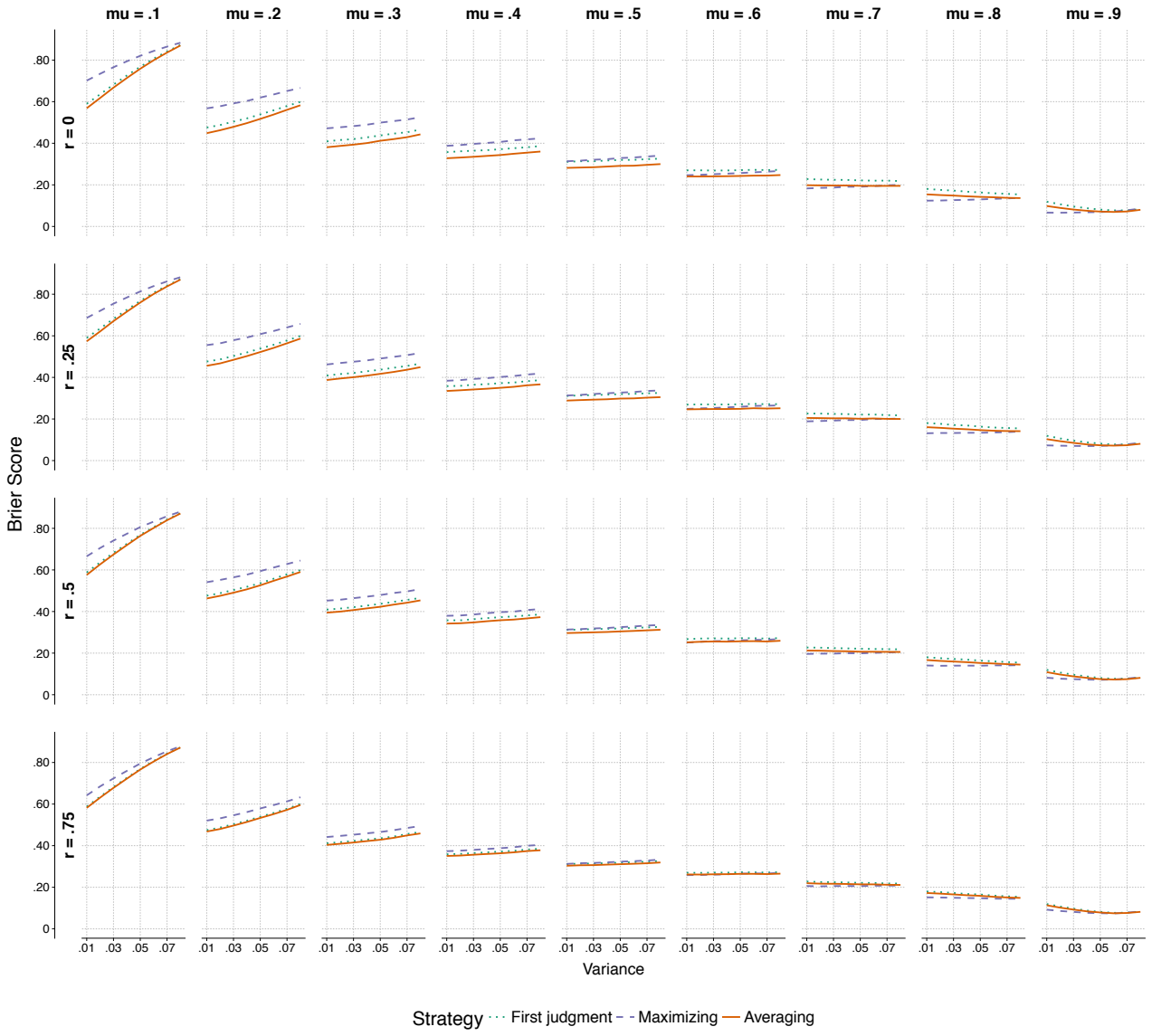


Figure A7. Brier scores (y-axes) of simulated strategies in heterogeneous environments varying in the mean of $p(C)$ (μ; columns), r (rows), and the variance of $p(C)$ (σ^2; x-axes). The correlation values r [0, .25, .5, .75] represent increasingly dependent knowledge sources underlying the two confidence judgments.

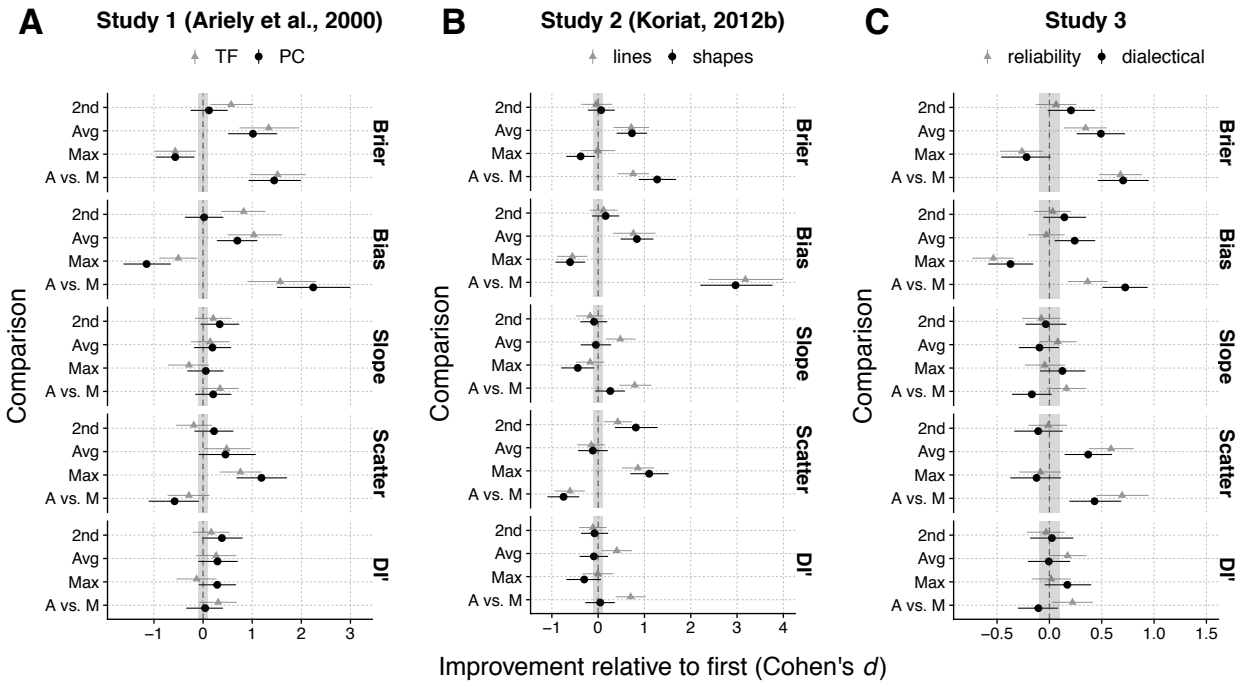


Figure A8. Effect sizes (Cohen's d) of differences in Brier decomposition measures (Brier score, bias, slope, scatter, and DI') between first minus second (2nd), first minus averaged (Avg), first minus maximized (Max), and maximized minus averaged (A vs. M) confidence judgments. We summarize the posterior distributions by reporting medians as point estimates and 95% highest density intervals (HDIs) as uncertainty intervals. Bars to the right of zero imply improved scores, and bars to the left of zero imply harmed scores. The shaded region ranges between -0.1 and 0.1, and constitutes the region of practical equivalence around the null value. PC = pairwise comparison condition; TF = true-or-false condition.

be reflected in the overall accuracy of maximizing, most likely because they are cancelled out by the negative effects of maximizing on bias scores. Consistent with the findings from our simulation study, maximizing increased bias scores throughout all studies (Figure A8, Table A5), whereas averaging reduced bias scores throughout all studies (except in the reliability condition of Study 3).

A5 Additional Results on Participants' Behavior

Results across studies and conditions show that participants changed their decisions, on average, in 12% to 22% of questions (Table A6). Furthermore, we calculated the mean confidence level of participants' initial decisions, separately for changed and unchanged decisions (Table A7), thus being able to examine if changing one's decision was related to initial confidence level. Results across conditions and studies show that confidence in the initial decision was lower when participants changed their decision later on as compared to when they did not. The mean within participant differences (Δ_M) are larger than their standard deviations (Δ_{SD}) in all studies and conditions, suggesting substantial effect sizes. This observation is in line with models of confidence, such as the *self-consistency model* (Koriat, 2012a) or the *Two-stage Signal Detection Model* (Pleskac & Busemeyer, 2010).

Table A2. Effect Sizes (Cohen's d) of Differences in Slope Between First Versus Averaged and First Versus Maximized Confidence Judgments

Study	Condition	Cohen's *d*	95%HDI
First vs. averaging			
Study 1 Ariely et al. (2000)	PC	0.189	[-0.183, 0.571]
	TF	0.147	[-0.250, 0.542]
Study 2 Koriat (2012b)	Shapes	-0.047	[-0.378, 0.279]
	Lines	0.476	[0.160, 0.808]
Study 3	Dialectical	-0.095	[-0.292, 0.091]
	Reliability	0.080	[-0.095, 0.260]
First vs. maximizing			
Study 1 Ariely et al. (2000)	PC	0.054	[-0.320, 0.414]
	TF	-0.275	[-0.717, 0.122]
Study 2 Koriat (2012b)	Shapes	-0.433	[-0.799, -0.087]
	Lines	-0.174	[-0.479, 0.137]
Study 3	Dialectical	0.122	[-0.090, 0.344]
	Reliability	-0.046	[-0.235, 0.150]

Note. HDI = Highest density interval. PC = pairwise comparison; TF = true or false. Cohen's d = median value of the posterior distribution; 95% HDI = 95% highest density interval of the posterior distribution.

Table A3. Effect sizes (Cohen's d) of differences in DI' Scores Between First Versus Averaged and First Versus Maximized Confidence Judgments

Study	Condition	Cohen's *d*	95%HDI
First vs. averaging			
Study 1 Ariely et al. (2000)	PC	0.289	[-0.096, 0.706]
	TF	0.266	[-0.147, 0.678]
Study 2 Koriat (2012b)	Shapes	-0.090	[-0.400, 0.214]
	Lines	0.398	[0.077, 0.731]
Study 3	Dialectical	-0.006	[-0.205, 0.198]
	Reliability	0.173	[-0.005, 0.354]
First vs. maximizing			
Study 1 Ariely et al. (2000)	PC	0.287	[-0.090, 0.673]
	TF	-0.125	[-0.542, 0.273]
Study 2 Koriat (2012b)	Shapes	-0.288	[-0.686, 0.060]
	Lines	-0.012	[-0.342, 0.327]
Study 3	Dialectical	0.169	[-0.046, 0.400]
	Reliability	0.015	[-0.170, 0.206]

Note. HDI = Highest density interval. PC = pairwise comparison; TF = true or false. Cohen's d = median value of the posterior distribution; 95% HDI = 95% highest density interval of the posterior distribution.

Table A4. Effect sizes (Cohen's d) of differences in Scatter Between First Versus Averaged and First Versus Maximized Confidence Judgments

Study	Condition	Cohen's d	95%HDI
First vs. averaging			
Study 1 Ariely et al. (2000)	PC	0.429	[-0.088, 1.072]
	TF	0.266	[-0.147, 0.678]
Study 2 Koriat (2012b)	Shapes	0.471	[0.015, 0.965]
	Lines	-0.150	[-0.457, 0.154]
Study 3	Dialectical	0.370	[0.146, 0.598]
	Reliability	0.587	[0.377, 0.806]
First vs. maximizing			
Study 1 Ariely et al. (2000)	PC	1.178	[0.680, 1.710]
	TF	0.756	[0.339, 1.190]
Study 2 Koriat (2012b)	Shapes	1.094	[0.694, 1520]
	Lines	0.855	[0.512, 1.208]
Study 3	Dialectical	-0.120	[-0.374, 0.111]
	Reliability	-0.085	[-0.289, 0.110]

Note. HDI = Highest density interval. PC = pairwise comparison; TF = true or false. Cohen's d = median value of the posterior distribution; 95% HDI = 95% highest density interval of the posterior distribution.

Table A5. Effect sizes (Cohen's d) of Differences in Bias Between First Versus Averaged and First Versus Maximized Confidence Judgments

Study	Condition	Cohen's d	95%HDI
First vs. averaging			
Study 1 Ariely et al. (2000)	PC	0.696	[0.281, 1.105]
	TF	1.014	[0.499, 1.604]
Study 2 Koriat (2012b)	Shapes	0.833	[0.487, 1.194]
	Lines	0.737	[0.326, 1.241]
Study 3	Dialectical	0.241	[0.050, 0.437]
	Reliability	-0.029	[-0.205, 0.146]
First vs. maximizing			
Study 1 Ariely et al. (2000)	PC	-1.145	[-1.618, -0.656]
	TF	-0.501	[-0.898, -0.115]
Study 2 Koriat (2012b)	Shapes	-0.604	[-0.923, -0.282]
	Lines	-0.554	[-0.886, -0.236]
Study 3	Dialectical	-0.371	[-0.586, -0.155]
	Reliability	-0.534	[-0.737, -0.342]

Note. HDI = Highest density interval. PC = pairwise comparison; TF = true or false. Cohen's d = median value of the posterior distribution; 95% HDI = 95% highest density interval of the posterior distribution.

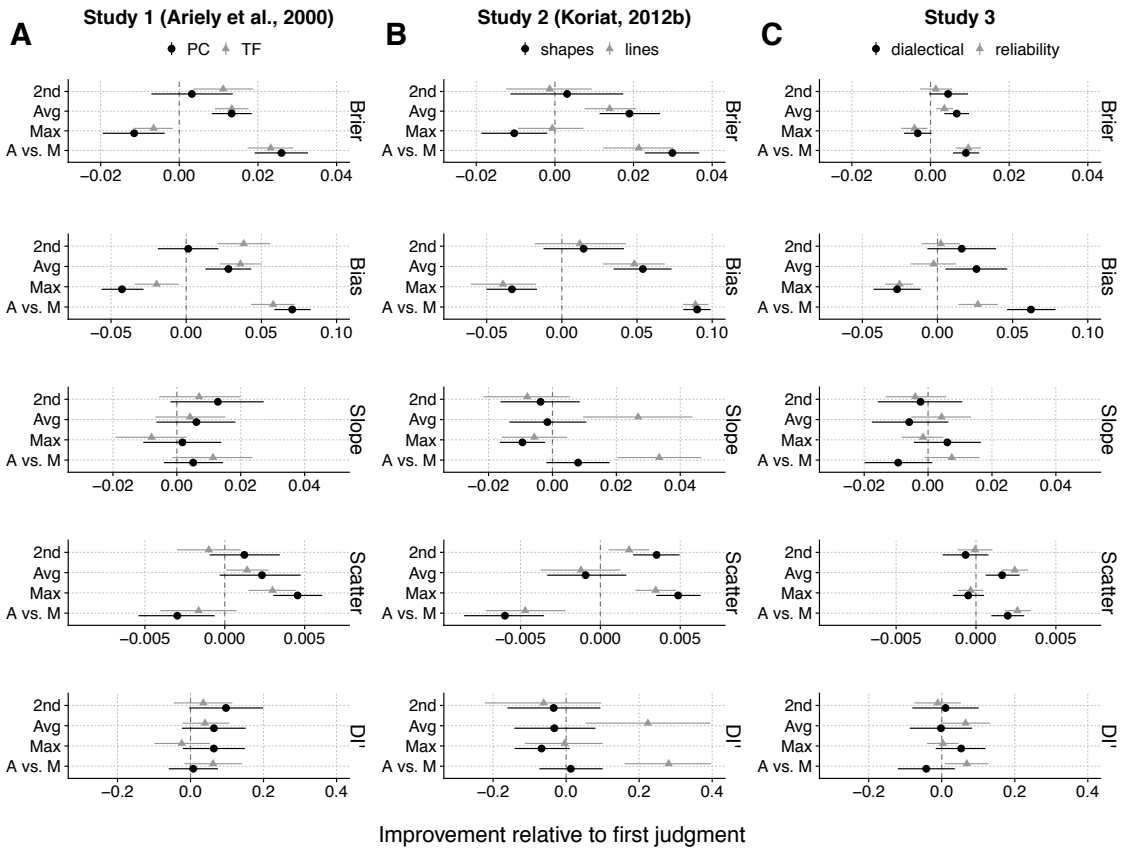


Figure A9. Mean differences in Brier decomposition measures between first minus second (2nd), first minus averaged (Avg), first minus maximized (Max), and maximized minus averaged (A vs. M) confidence judgments, by study and accuracy measure. We summarize the posterior distributions by reporting medians as point estimates and 95% highest density intervals (HDIs) as uncertainty intervals. PC = pairwise comparison; TF = true or false..

Additionally, we calculated differences in accuracy (in terms of proportion of correct decisions) between first and second, first and averaged and first and maximized decisions (Table A8). In almost all studies, the 95%-HDI (highest density interval) includes the zero value, indicating no effect on the proportion of correct decisions. Study 3 constitutes the only exception, with second and averaged dialectical estimates showing a mean increase of 2 percentage points.

Finally, we calculated the difference between first and second confidence judgments in general, that is, irrespective of whether participants changed their decision in the second phase (Table A9). Participants' mean confidence for second judgments decreased slightly in Study 1 and 2, by 0.0015 and 0.015, respectively; we found no changes in Study 3.

Table A6. Proportion of changed decisions per study and condition.

Study	Median	IQR
Study 1 (Ariely et al., 2000)	0.22	[0.15, 0.35]
Study 2 (Koriat, 2012b)	0.21	[0.19, 0.25]
Study 3 (New Experiment)	0.20 (dialectical)	[0.10, 0.32]
	0.12 (reliability)	[0.04, 0.16]

Note. We report median and interquartile range (IQR) as the distributions are markedly skewed towards zero.

Table A7. Mean confidence of first decisions as a function of whether or not the second decision was different.

Study	Change	Mean	SD	Δ_M	Δ_{SD}
Study 1 (Ariely et al., 2000)	no	084	0.07	0.19	0.09
	yes	0.65	0.10		
Study 2 (Koriat, 2012b)	no	0.78	0.09	0.06	0.04
	yes	0.72	0.09		
Study 3 (New Experiment)	no	0.71(dialectical)	0.10	0.13	0.10
	yes	0.58 (dialectical)	0.08		
	no	0.67 (reliability)	0.08	0.12	0.10
	yes	0.55 (reliability)	0.09		

Note. We report the mean and standard deviation (SD) of mean confidence across participants; in addition, we report the mean of the within-participant differences in mean confidence for changed vs. non-changed decisions (Δ_M), where positive differences indicate that the changed decision was more confident. Furthermore, we report the standard deviation of those within differences (Δ_{SD}).

Table A8. Mean differences in proportion of correct decisions between first minus second, first minus averaged and first minus maximized decisions.

Study	Comparison	Condition	Δ_M	95% HDI
Study 1 (Ariely et al., 2000)	Second	PC	-0.015	[-0.031, 0.001]
		TF	0.007	[-0.006, 0.022]
	Averaging	PC	0.0	[-0.010, 0.010]
		TF	0.009	[-0.002, 0.020]
	Maximizing	PC	-0.005	[-0.016, 0.005]
		TF	0.009	[-0.002, 0.020]
Study 2 (Koriat, 2012b)	Second	Shapes	-0.003	[-0.025, 0.018]
		Lines	0.002	[-0.026, 0.020]
	Averaging	Shapes	0.008	[-0.006, 0.023]
		Lines	0.004	[-0.013, 0.021]
	Maximizing	Shapes	0.008	[-0.004, 0.023]
		Lines	0.002	[-0.015, 0.021]
Study 3 (New Experiment)	Second	Dialectical	0.025	[0.003, 0.046]
		Reliability	0.009	[-0.002, 0.020]
	Averaging	Dialectical	0.021	[0.001, 0.040]
		Reliability	-0.002	[-0.017, 0.012]
	Maximizing	Dialectical	0.008	[-0.005, 0.023]
		Reliability	0.005	[-0.002, 0.013]

Note. HDI = Highest density interval. PC = pairwise comparison; TF = true or false. Positive differences imply an improved proportion of correct decisions. PC = pairwise comparison; TF = true or false.

Table A9. Difference between first and second confidence judgments.

Study	Δ_M	95%-HDI
Study 1 (Ariely et al., 2000)	0.0015	[0.0007, 0.0023]
Study 2 (Koriat, 2012b)	0.015	[0.0110, 0.0190]
Study 3 (New Experiment)	-0.0008 (dialectical)	[-0.0015, 0.0001]
	0.00 (reliability)	[0.0000, 0.0000]

Note. HDI = Highest density interval. We report the mean (Δ_M) and 95%-HDI of within-participant differences in confidence between first and second judgments. Positive differences indicate that the second judgment was less confident.

References

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., ... Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*, 130–147. doi: 10.1037/1076-898X.6.2.130

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, *78*(1), 1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Koriat, A. (2012a). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. doi: 10.1037/a0025648

Koriat, A. (2012b). When are two heads better than one and why? *Science*, *336*(6079), 360–362. doi: 10.1126/science.1216549

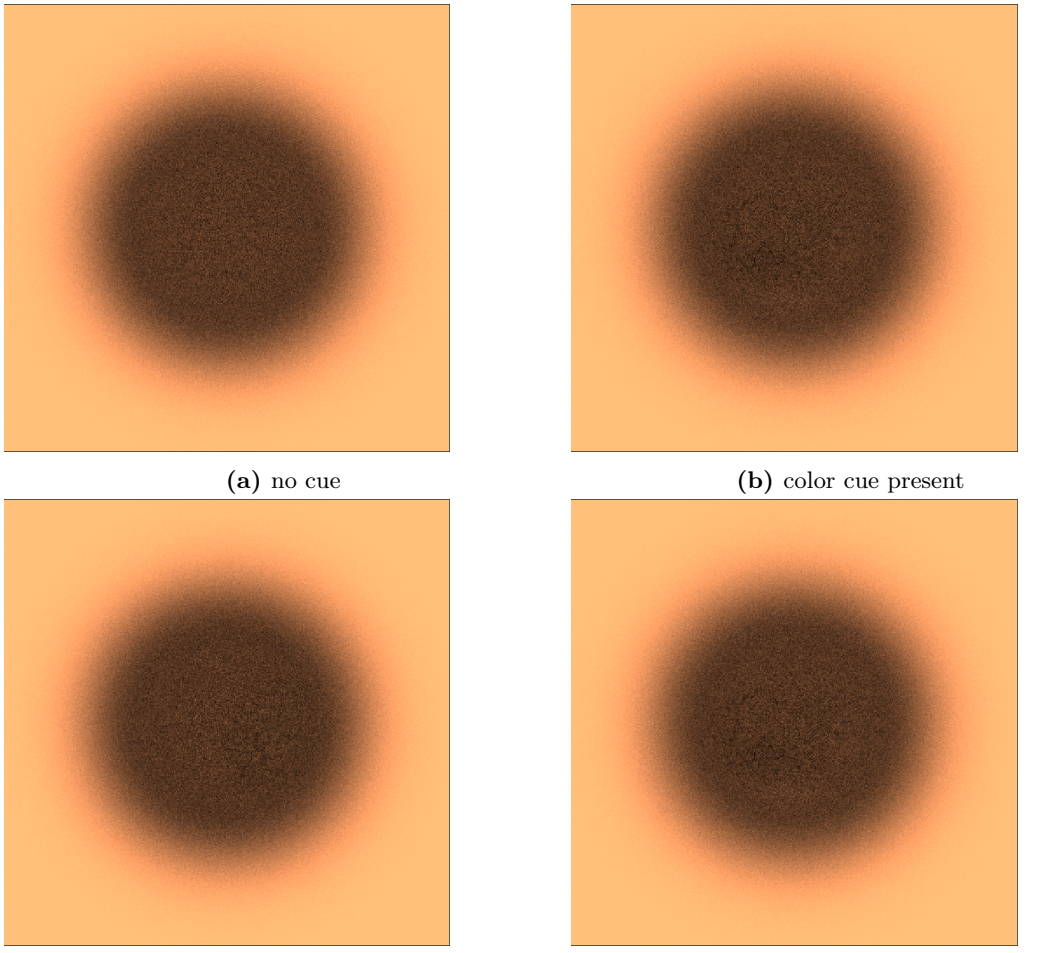
Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. doi: 10.1037/a0019737

Wallsten, T. S., & Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, *41*(1), 1–18. doi: 10.1016/S0165-4896(00)00053-6

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice-Hall.

B | Supplementary Material to Chapter 4: "Cognitive Dependencies in Sequential Diagnostic Reasoning Tasks"

B1 Example Stimuli Used in the Experiment



(a) no cue

(b) color cue present

(c) structure cue present

(d) color and structure cues present

Figure B1. Example stimuli

B2 Signal Detection Model in JAGS

```
model{
        # Hyperpriors
        for (e in 1:nenv) {
                for (c in 1:ncond) {
                        lambdac[e,c] ~ dgamma(.001,.001)
                        lambdad[e,c] ~ dgamma(.001,.001)
                        sigmac[e,c] <- 1/sqrt(lambdac[e,c])
                        sigmad[e,c] <- 1/sqrt(lambdad[e,c])

                        muc[e,c] ~ dnorm(0,.001)
                        mud[e,c] ~ dnorm(0,.001)
                        mucPrior[e,c] ~ dnorm(0,.001)
                        mudPrior[e,c] ~ dnorm(0,.001)
                        }
                }

        # Priors
        for (s in nSub) {
                for (cc in 1:ncond) {
                        c[s,cc] ~ dnorm(muc[env[s],cc],lambdac[env[s],cc])
                        d[s,cc] ~ dnorm(mud[env[s],cc],lambdad[env[s],cc])

                        # reparameterization using equal variance gaussian SDT
                        thetah[s,cc] <- phi(d[s,cc]/2-c[s,cc])
                        thetaf[s,cc] <- phi(-d[s,cc]/2-c[s,cc])

                        # observed data
                        h[s,cc] ~ dbin( thetah[s,cc], sgnl[s,cc])
                        fa[s,cc] ~ dbin( thetaf[s,cc], noise[s,cc])

                        # posterior prediction
                        hitPost[s,cc] ~ dbin(thetah[s,cc], sgnl[s,cc])
                        faPost[s,cc] ~ dbin(thetaf[s,cc], noise[s,cc])
                }
        }
}
```

B3 Multimodal Response Time Distributions

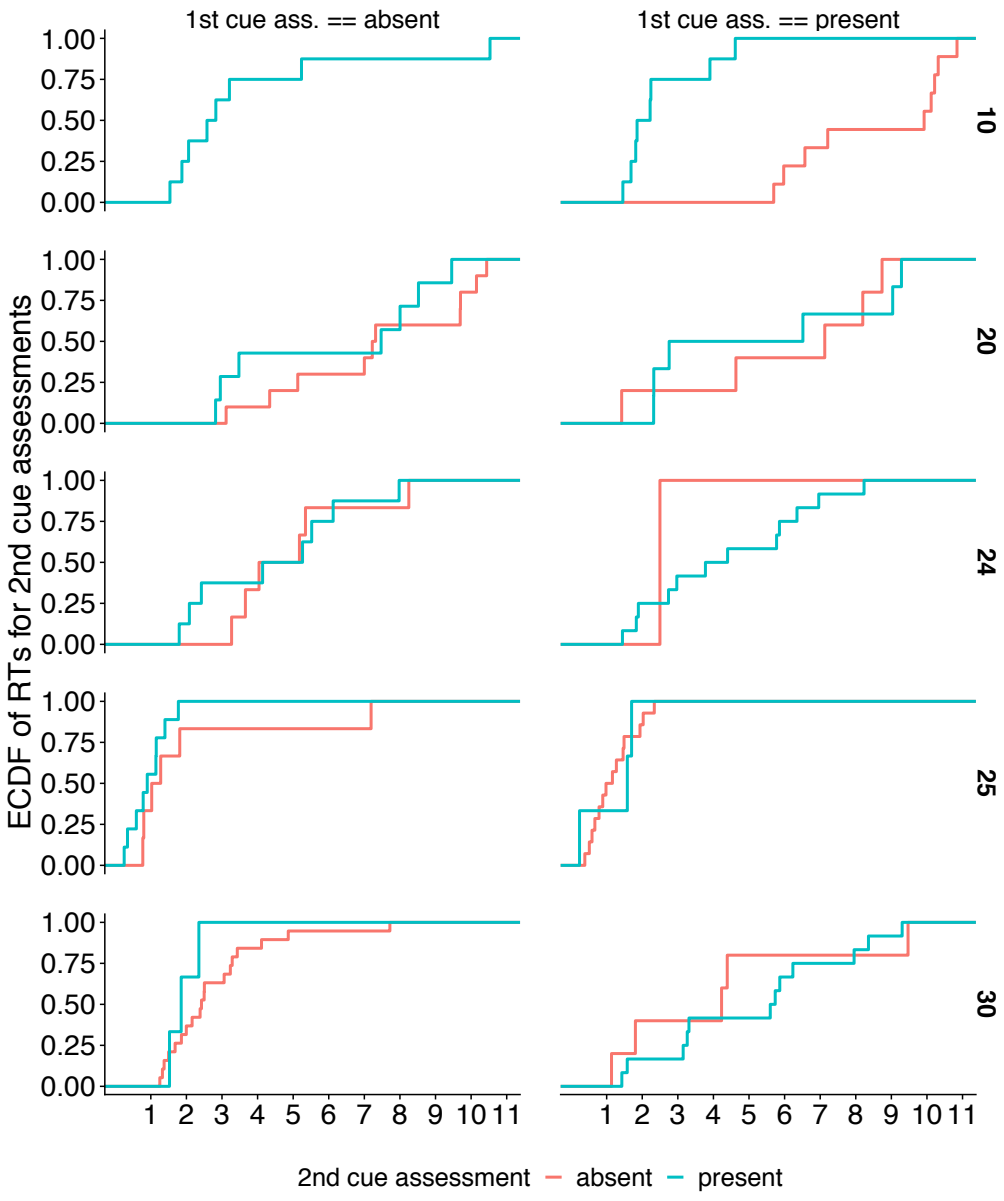


Figure B2. Reaction Times. Empirical cumulative distribution function (ECDF, y-axis) of participants' reaction times (x-axis) in the testing phase, separately per decision on first assessment (columns) and decision on second assessment (color) for subjects 10, 20, 24, 25, and 30. The ECDF depicts the relative proportion of each observed value ordered on the x-axis. The dotted horizontal line corresponds to 50% of observed values and therefore indicates the median. The minimum and maximum of the distribution is illustrated by the begin (on the x-axis) and the end (y-axis) of the line. The steeper the curve, the narrower the distribution. Plateaus indicate multimodality.

Acknowledgments

For reasons of data protection, the acknowledgements are not included in the online version.

Curriculum Vitae

ALEKSANDRA LITVINOVA, M.Sc.

Max Planck Institute for Human Development Center for Adaptive Rationality (ARC)
Lentzeallee 94, 14195 Berlin, Germany *Phone:* +49–(0)30–824–06–598
Email: litvinova@mpib-berlin.mpg.de

EDUCATION

2015 – 2018 **Doctoral Studies in Psychology**
Freie Universität Berlin & MPI for Human Development, Center for Adaptive Rationality, Germany
PIs: Stefan M. Herzog & Ralph Hertwig, Advisor: Stefan M. Herzog

2012 – 2014 **Research Master in Neuroeconomics**
Maastricht University, The Netherlands
Master Thesis: Dynamics of Moral Negotiations

2009 – 2012 **Bachelor in Psychology**
Maastricht University, The Netherlands
Bachelor Thesis: Consciousness in Split-Brain Patients
Fall 2011: Semester abroad at City University, Hong Kong

RESEARCH

2014 – 2015 **Project Specialist**
Computational Social Science Laboratory, University of Southern California
Project 1: Modeling and exploiting the social function of emotions in mixed human-machine teams; PI & advisor Morteza Dehghani
Project 2: Investigating & Modeling the Emergence of Counterfactual Reasoning; PI: Henrike Moll & Morteza Dehghani; advisor: Morteza Dehghani

2013 – 2014 **Visiting Scholar**
Brain and Creativity Institute, University of Southern California
Project: Modeling and exploiting the social function of emotions in mixed human-machine teams; PI & Advisor: Morteza Dehghani

PUBLICATIONS

Herzog, S. M., **Litvinova, A.**, Yahosseini, K. S., Tump, A. N., & Kurvers, R. K. J. M. (2019). The ecological rationality of the wisdom of crowds. In R. Hertwig, T. J. Pleskac, T. Pachur, & The Center for Adaptive Rationality (Eds.), *Taming uncertainty* (pp. 245–262). Cambridge, MA: MIT Press.

Kim, E., Gimbel, S., **Litvinova, A.**, Kaplan, J., Dehghani, M. (2017). Decoding Virtual Agent's Emotion and Strategy from Brain Patterns. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2407–2412). Austin, TX: Cognitive Science.

Kim, E., Gimbel, S., **Litvinova, A.**, Kaplan, J., Dehghani, M. (2016). Predicting Decision in Human-Agent Negotiation using functional MRI. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.) (2016). *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 638–643). Austin, TX: Cognitive Science Society.

Boghrati, R., Garten, J., **Litvinova, A.**, Dehghani, M. (2015). Incorporating Background Knowledge into Text Classification. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 244–249). Austin, TX: Cognitive Science Society.

In preparation and submitted

Litvinova, A., Herzog, S. M., Kall, A. A., Pleskac, T. J., & Hertwig, R. (in press). How the "Wisdom of the Inner Crowd" Can Boost Accuracy of Confidence Judgments. Decision.

Litvinova, A., Herzog, S. M., Hertwig, R., de Zoete, A., Ostelo, R., & Kurvers, R.H.J.M. (in prep). When do Experts Change Their Mind?

Litvinova, A., Herzog, S. M., Kurvers, R.H.J.M., & Hertwig, R. (in prep). Cognitive dependencies in sequential diagnostic reasoning tasks.

Moll, H., Pettit, C., **Litvinova, A.**, Min, J., & Dehghani, M. (submitted). Do Preschoolers Reason Counterfactually When Spontaneously Revising Choices? *Cognition and Emotion.*

CONFERENCE CONTRIBUTIONS

Talks (presenter in bold)

Litvinova, A., Herzog, S. M., Pleskac, T. J., & Hertwig, R. (2017). *Harnessing the Wisdom of the Inner Crowd by Exploiting Confidence.* Talk presented at the 26th Subjective Probability, Utility and Decision Making Conference (SPUDM), Haifa, Israel.

Litvinova, A., Herzog, S. M., & Hertwig, R. (2016). *Harnessing the wisdom of the inner crowd by exploiting the confidence in your decisions.* Talk presented at the JDMx 2016 meeting, Fakultät für Psychologie, Universität Basel, Switzerland.

Posters

Litvinova, A., Herzog, S. M., Kurvers, R.H.J.M., & Hertwig, R. (2017). *When do experts change their mind?* Poster presented at the Scientific Advisory Board Meeting, Berlin, Germany.

Litvinova, A., Herzog, S. M., & Hertwig, R. (2016). *Harnessing the wisdom of the inner crowd by exploiting the confidence in your decisions.* Poster presented at the 37th Annual Meeting of the Society for Judgment and Decisions Making, Boston, Massachusetts, USA.

Litvinova, A., Herzog, S. M., & Hertwig, R. (2016). *Harnessing the wisdom of the inner crowd by exploiting the confidence in your decisions.* Poster presented at the 57th Annual Meeting of the Psychonomics Society, Boston, Massachusetts, USA.

Declaration of Independent Work

I hereby declare that I completed the doctoral thesis independently. Except where otherwise stated, I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution. I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Department of Education and Psychology of Freie Universität Berlin, as amended on 8th August 2016. The principles of Freie Universität Berlin for ensuring good academic practice have been complied with.

Aleksandra Litvinova

Berlin, 22 November 2018