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

Aleksandra Litvinova, Stefan M. Herzog

Max Planck Institute for Human Development, Center for Adaptive Rationality

Alice A. Kall

Recruiting Partner Helvetia Insurance Group

Timothy J. Pleskac, Ralph Hertwig

Max Planck Institute for Human Development, Center for Adaptive Rationality

## Author Note

Data and scripts for reproduction can be found at `https://osf.io/b3f6d/`. 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).

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. However, what if there are no other individuals' judgments at one's disposal? We investigated when an individual can avail theirself 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 two-alternative choice tasks. 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) 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—a result, which has not been established in prior work. 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

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

## 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 et al., 1982; Sniezek et al., 1990). Notwithstanding this notorious reputation (but see Gigerenzer et al., 1991; Juslin et al., 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 et al., 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; Moore et al., 2015; McClelland & Bolger, 1994). 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 (Gigerenzer et al., 1991; Juslin et al., 2000; Dhami et al., 2004). 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 et al., 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 et al., 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*. The focus of our study is prescriptive, that is, we will investigate in which environments which aggregation strategies for confidence make sense irrespective of how the strategies will be ultimately implemented—informally by humans or mechanically by software.

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 (Herzog et al., 2019; Larrick et al., 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 et al., 2009; Steyvers et al., 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 (Vul & Pashler, 2008; Van Dolder & van den Assem, 2018; but see Steegen et al., 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, which can be interpreted as the central tendency of their subjective probability distribution. However, no attention was paid to individual's uncertainty in their estimates and how aggregation changes a person's uncertainty or confidence. The main principle behind aggregation gains for quantitative estimates is bracketing: if two or more values bracket the true value averaging can reduce

error. However, for categorical decisions with an associated confidence judgment, the bracketing principle does not apply. 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 et al., 1991; Yu et al., 2015). As a result, confidence can serve as a cue to the accuracy of a decision or forecast (Bang et al., 2014; Kämmer et al., 2017; Koriat, 2012b; Kurvers et al., 2016). 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 et al., 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, selecting the highest confidence judgment is the most advantageous. Consider the extreme case of a decision maker who is always correct, but who does not always report 100% confidence. Maximizing confidence judgments will improve diagnosticity by moving the final confidence judgments closer to 100%. However, if confidence tracks consensuality and not accuracy per se, as suggested by Koriat (2012a), then 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 et al., 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 & McGoey (1980, their 2AFC(HR) model) and further developed by Gu & 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" (Macmillan & Creelman, 2004) for first and second confidence judgments, we drew *one* value from either the signal or the noise distribution. An individual's subjective intensity reflects the overall evidence accumulated, in other words, it "embod[ies] all the information the respondent has about whether the answer is true or the response is right" (Ferrell & McGoey, 1980, p. 34). Whether the value was drawn from the signal or the noise distribution was determined by drawing either 1 or 0 from a Bernoulli distribution where

[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 supplementary material, subsection "Heterogeneous Environments").

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 $\Phi$ 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).

Finally, to translate the subjective, latent intensities into overt confidence judgments, we followed Ferrell & McGoey (1980) and Gu & 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

[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 et al., 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.

probability scale. Furthermore, the model assumes that confidence judgments in choice tasks with two mutually exclusive alternatives add up to 1 (additivity). Thus, judgments that fell below 50% thus imply that the belief in being correct was higher for the opposite decision. Previous research has shown, that individuals' actual confidence judgments of complementary statements show almost perfect additivity, with minor deviations of only .015 points from 1 (sd = .033; Wallsten et al., 1993; Ariely et al., 2000).

## 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 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 supplementary material, section "Decomposition of Overall Accuracy in the Simulation"). To summarize, the effects of averaging and maximizing depend simultaneously on the mean ($\mu$) and variance of $p(C)$ of the environment. In general, as $\mu$ increased, the Brier score of all strategies improved. The effect of variance on the performance of confidence judgments depends on $\mu$: 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 A 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

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

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 supplementary material, section "Decomposition of Overall Accuracy in the Simulation"). 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.

**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 $p(C) > 0.6$ and not already once $P(C) > 0.5$, as one might have intuited based on the notion that maximizing only works for kind items. That is, a kind item is a necessary but not a sufficient condition for maximizing to outperform averaging. In our reading of the literature, this is a result, which has not been established in prior work. 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 & 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 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. Assuming that confidence judgments in choice tasks with mutually exclusive alternatives add up to 1 (additivity; Wallsten et al., 1993; Ariely et al., 2000), we aggregated confidence judgments of opposing decisions by converting confidence in the given answer to confidence in the correct answer and then calculated the average confidence.

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 & Hertwig (2009) first tested the dialectical bootstrapping approach in a quantitative estimation task using the consider-the-opposite technique (adapted from Lord et al., 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

[4]On the basis of the medium effect of the dialectical instruction on the accuracy of quantitative estimates observed in Herzog & 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.

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 B, Table B1). 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 et al., 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.

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 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 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 evidence that averaging dialectical judgments enhanced the Brier score more than merely averaging reliability judgments ($d = 0.28$, 95% HDI [-0.02, 0.59], with a

posterior probability of 88% that the effect size of the difference is relevant (i.e., $d > +0.1$).

***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 supplementary material, section "Decomposition of Overall Accuracy in the Empirical Studies" 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's accuracy of confidence judgment can be enhanced 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. In practice, these strategies could be implemented either by an individual herself or mechanically by a software. For example, consider a radiologist who evaluates x-rays and registers her diagnosis and confidence judgment on a computer. A software could average or maximize a radiologist's two confidence ratings across two sessions, which could be separated by, say, several hours. 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 et al., 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 &

[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.

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 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 et al., 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

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

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

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

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

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

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., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, *39*(2), 176–190.

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

**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.*

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.*

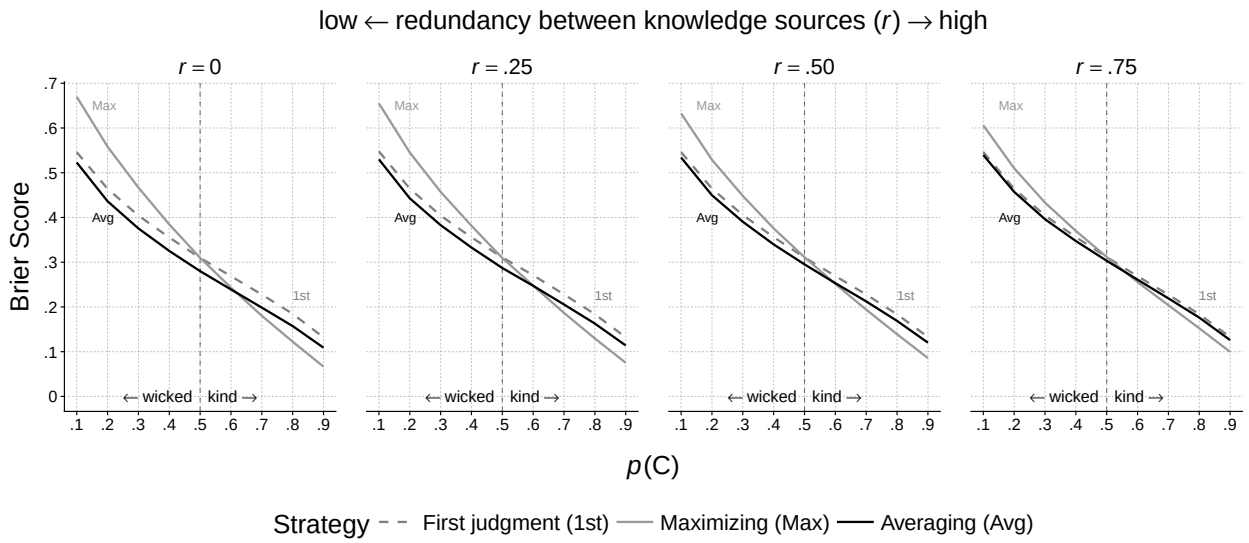


*Figure 1*. Overall accuracy of simulated confidence judgments as measured by the Brier score ($y$ axis), where lower values indicate better quality. Panels (from left to right) correspond to increasingly more redundant knowledge sources underlying the two confidence judgments (correlation values $r$). 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.

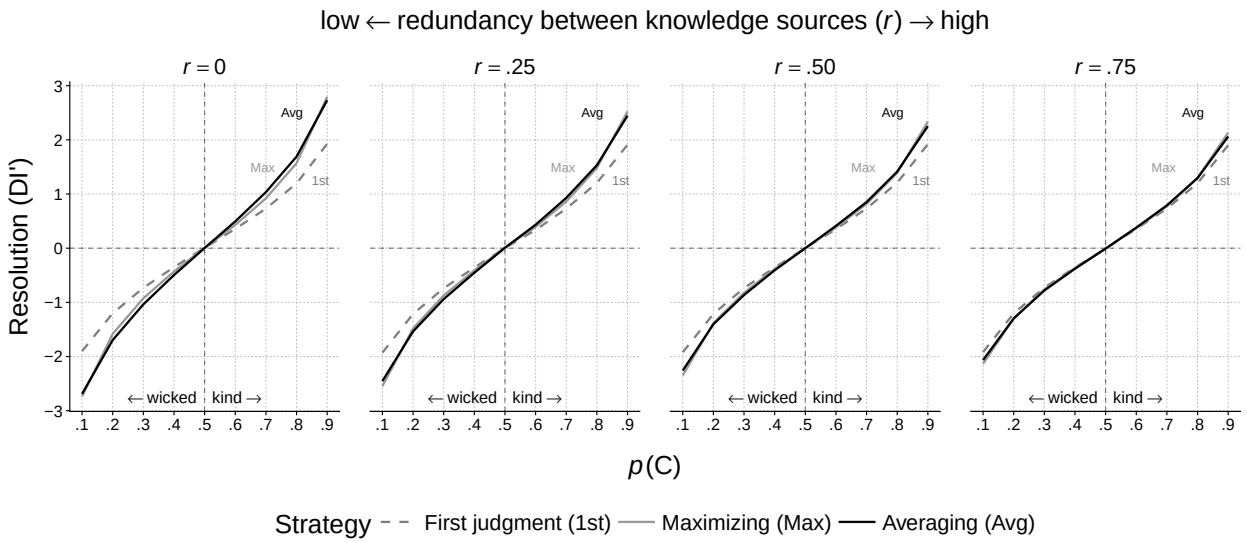


*Figure 2*. Resolution of simulated confidence judgments as measured by $DI'$ ($y$ axis). $DI'$ quantifies the ability of confidence judgments to discriminate between correct and wrong decisions (i.e., difference between mean confidence of correct vs. incorrect decisions, standardized by the pooled $SD$ 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.

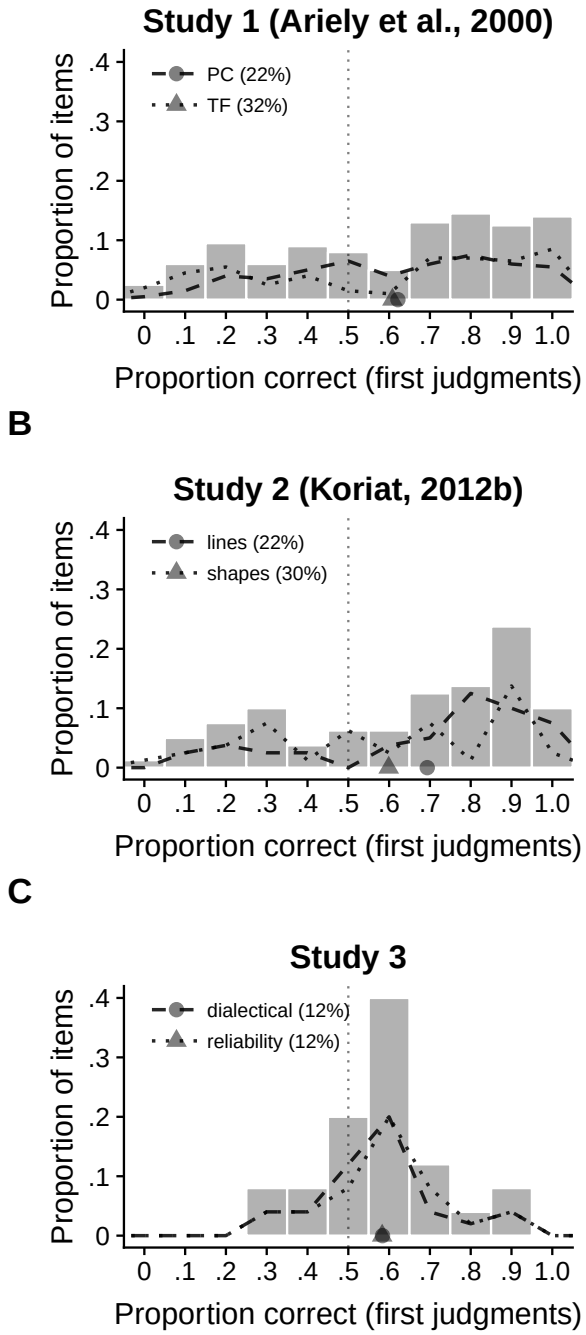


*Figure 3*. Histograms of proportion correct of items (based on first judgments), separately for each study. Dashed and dotted lines show the distributions per condition (A, C) or task (B). Circles and triangles on the bottom of each panel indicate median proportion correct across items per condition (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.

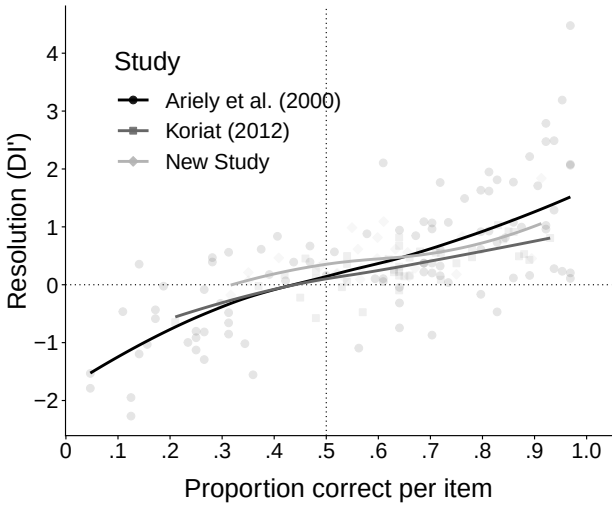


*Figure 4*. Confidence's ability to distinguish between correct and wrong first answers per item ($DI'$ per item; $y$ axis) as a function of proportion correct of that item ($x$ axis). Results are shown separately for each study (pooled across the two tasks in Study 1 (Ariely et al., 2000) and Study 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*).

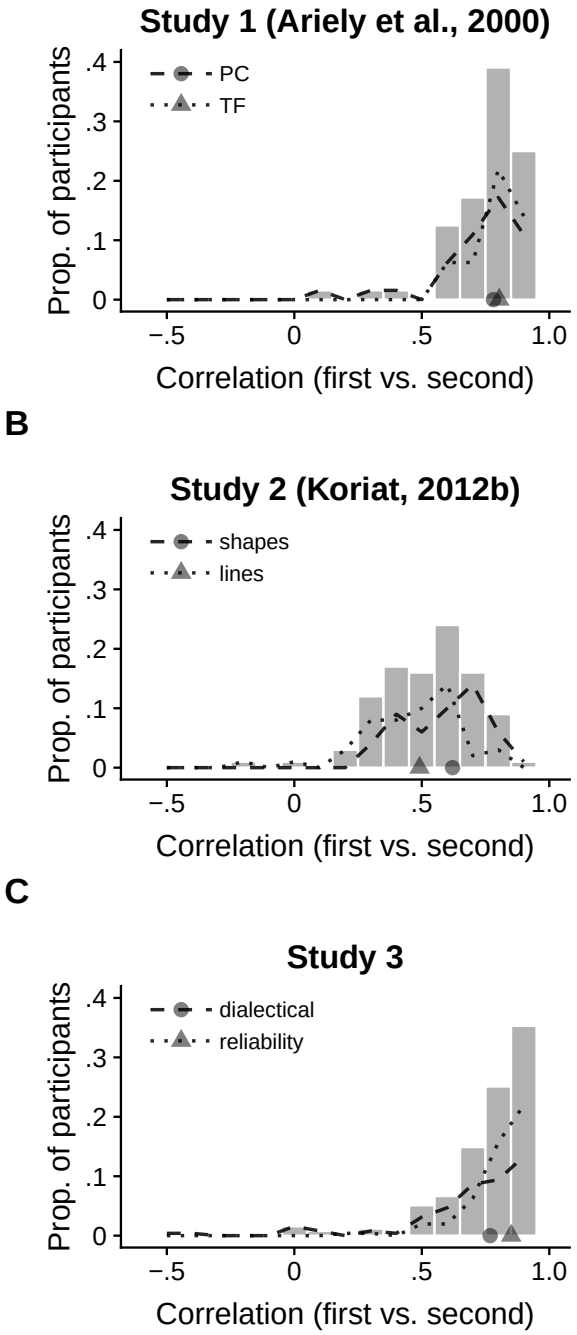


*Figure 5*. Histogram of Spearman correlations between first and second confidence judgments separately for study. Dashed and dotted lines show distributions per condition (A, C) or task (B). Circles and triangles on the bottom of each panel indicate median correlations per condition or task. Study 1 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.

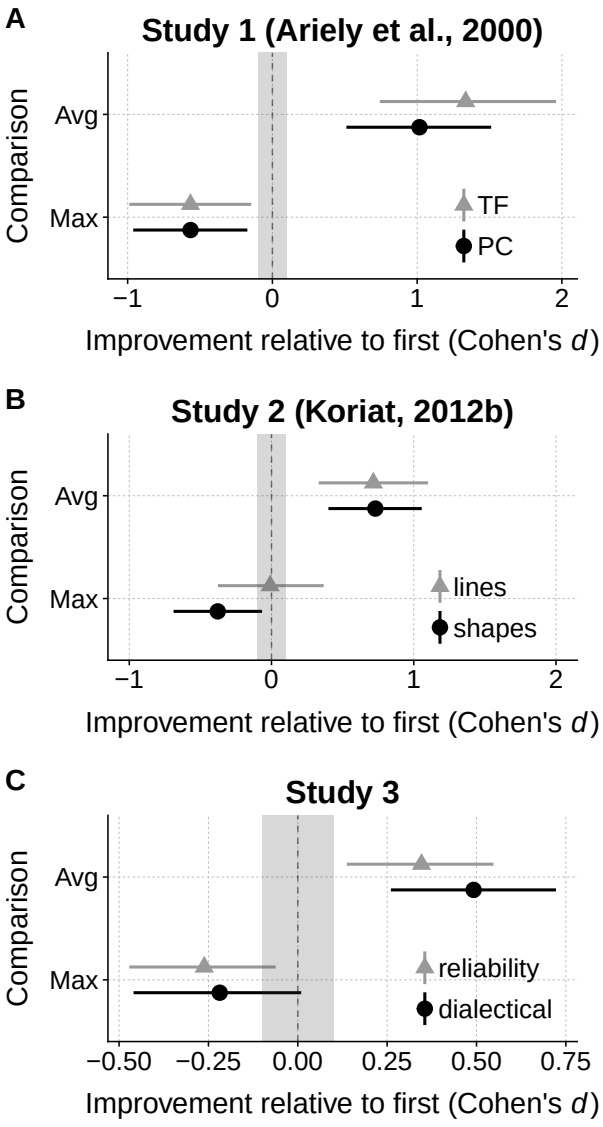


*Figure 6*. Effects of averaging (Avg) and maximizing (Max) on the overall accuracy of confidence judgments relative to first judgments. The $x$ axis shows the improvement in Brier scores expressed as Cohen's $d$ effect sizes; symbols show the median value and the ranges show the 95% highest density 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.

## Appendix A

### 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.

# Appendix B

## Items Used in Study 3

Table B1
*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.*