Journal of Memory and Language 71 (2014) 17–38

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Knowing the crowd within: Metacognitive limits on combining multiple judgments

Scott H. Fraundorf*, Aaron S. Benjamin

University of Illinois at Urbana-Champaign, United States

ARTICLE INFO

Article history:
Received 8 January 2013
revision received 27 August 2013
Available online 13 November 2013

Keywords:
Metacognition
Subjective fluency
Judgment and decision making

ABSTRACT

We investigated how decision-makers use multiple opportunities to judge a quantity. Decision-makers undervalue the benefit of combining their own judgment with an advisor's, but theories disagree about whether this bias would apply to combining several of one's own judgments. Participants estimated percentage answers to general knowledge questions (e.g., *What percent of the world's population uses the Internet?*) on two occasions. In a final decision phase, they selected their first, second, or average estimate to report for each question. We manipulated the cues available for this final decision. Given cues to general theories (the labels *first guess*, *second guess*, *average*), participants mostly averaged, but no more frequently on trials where the average was most accurate. Given item-specific cues (numerical values of the options), metacognitive accuracy was at chance. Given both cues, participants mostly averaged and switched strategies based on whichever yielded the most accurate value on a given trial. These results indicate that underappreciation of averaging estimates does not stem only from social differences between the self and an advisor and that combining general and item-specific cues benefits metacognition.

Introduction

The opportunity to revise a judgment offers both opportunity and challenge. Altering a business projection, reconsidering the accuracy of world knowledge retrieved from memory, or reassessing the time needed to complete a project affords the use of additional information not included in the original judgment. Indeed, making multiple estimates permits greater accuracy in judgment than what could be achieved with a single estimate: the aggregate of multiple estimates, even from the same individual, can outperform any single judgment by reducing the influence of random error on the judgment process (Herzog & Hertwig, 2009; Vul & Pashler, 2008), as detailed below.

However, a judge who has made multiple estimates also faces a decision about how to use those estimates: Is a particular estimate the most accurate; if so, which? Would the estimates be even better if aggregated? Although combining several estimates is generally the most effective strategy (Rauhut & Lorenz, 2010; Vul & Pashler, 2008), the literature suggests that decision-makers often do not make optimal use of multiple estimates. When given the opportunity to choose their own judgment, choose a judgment made by another person, or combine them, judges typically overrely on their own estimates even when judgment accuracy could be improved by combining them (Bonaccio & Dalal, 2006).

Using multiple self-generated estimates does not necessarily present the same challenges as estimates from other judges. One hypothesis is that the bias against combining one's own estimation with others' is due to social factors such as norms on how much advice should be taken or a belief that one is better than the average judge (Harvey & Fischer, 1997). This account does not predict similar underuse of averaging multiple estimates that are all self-generated and do not involve another person. An alternate hypothesis, however, is that suboptimal use of multiple judgments reflects broader cognitive chal-

* Corresponding author. Address: Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, United States.
E-mail address: scottfraundorf@gmail.com (S.H. Fraundorf).

0749-596X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jml.2013.10.002

Table 1
Across all studies, stimulus questions, correct answers, and mean and standard deviation of participant guesses for all trials; for trials in which the first estimate, second estimate, and average constituted three distinct integer values (75% of all trials); and for trials in which not all response options were distinct integer values (25% of all trials).

	Answer	Estimate 1		Estimate 2		Average	
		M	*SD*	*M*	*SD*	*M*	*SD*
All trials							
Q1: The area of the USA is what percent of the area of the Pacific Ocean?[a]	6.3	28.7	19.6	28.0	19.3	28.4	18.5
Q2: What percent of the world's population lives in either China, India, or the European Union?[a]	44.4	58.0	17.0	61.3	17.7	59.6	15.7
Q3: What percent of the world's airports are in the United States?[a]	30.3	33.4	20.4	34.1	19.7	33.8	18.2
Q4: What percent of the world's roads are in India?[a]	10.5	14.7	14.8	18.3	16.4	16.5	14.2
Q5: What percent of the world's countries have a higher fertility rate than the United States?[a]	58.0	36.4	23.1	37.2	24.6	36.8	21.8
Q6: What percent of the world's telephone lines are in China, the USA, or the European Union?[a]	68.0	72.1	17.9	64.4	21.3	68.2	16.6
Q7: Saudi Arabia consumes what percentage of the oil it produces?[a]	18.9	21.5	19.9	20.0	21.4	20.8	19.9
Q8: What percentage of the world's countries have a higher life expectancy than the United States?[a]	20.3	24.4	19.9	26.1	19.2	25.3	17.9
Q9: What percent of the United States population lives in Florida?	6.0	10.0	7.7	11.6	11.0	10.8	8.7
Q10: What percent of the world's population is 14 years of age or younger?	26.3	32.7	15.2	32.6	17.5	32.6	15.4
Q11: The Internet is used by what percent of the world's population?	30.3	60.6	23.1	58.3	25.0	59.5	23.3
Q12: The European Union consumes what percent of the world's electricity?	16.2	30.2	14.9	33.2	18.4	31.7	14.7
Trials with different first estimate, second estimate, and average (retained for analysis)							
Q1[a]	6.3	29.9	19.1	28.9	18.8	29.4	17.5
Q2[a]	44.4	57.8	17.1	61.8	17.8	59.8	15.4
Q3[a]	30.3	34.0	20.2	35.0	19.3	34.5	17.3
Q4[a]	10.5	16.0	15.6	21.0	17.2	18.5	14.6
Q5[a]	58.0	38.1	23.2	39.0	25.1	38.5	21.7
Q6[a]	68.0	70.8	18.8	61.3	21.9	66.1	16.8
Q7[a]	18.9	23.5	20.9	21.4	23.0	22.5	20.9
Q8[a]	20.3	25.7	20.7	27.9	19.7	26.8	18.2
Q9	6.0	11.2	8.2	13.8	12.4	12.5	9.4
Q10	26.3	33.0	15.2	32.9	18.3	33.0	15.4
Q11	30.3	60.9	22.7	57.8	25.1	59.3	22.9
Q12	16.2	29.8	14.3	33.3	18.5	31.6	14.1
Trials without different first estimate, second estimate, and average (excluded)							
Q1[a]	6.3	26.0	20.6	25.9	20.6	26.0	20.6
Q2[a]	44.4	58.9	17.2	59.0	17.2	59.0	17.2
Q3[a]	30.3	31.4	21.0	31.4	21.0	31.4	21.0
Q4[a]	10.5	11.4	12.0	11.4	12.0	11.4	12.0
Q5[a]	58.0	28.9	20.8	29.0	20.7	29.0	20.8
Q6[a]	68.0	77.7	11.4	77.8	11.5	77.8	11.4
Q7[a]	18.9	17.2	17.0	17.1	17.0	17.1	17.0
Q8[a]	20.3	20.1	16.0	20.1	16.0	20.1	16.0
Q9	6.0	7.7	6.3	7.7	6.2	7.7	6.2
Q10	26.3	31.7	15.4	31.7	15.5	31.7	15.5
Q11	30.3	59.9	24.7	60.0	24.8	59.9	24.7
Q12	16.2	32.4	17.8	32.5	17.8	32.5	17.8

Note: SD = standard deviation.
[a] Item used by Vul and Pashler (2008).

lenges—such as an incorrect belief about the mathematical value of averaging (Soll, 1999) or an overreliance on one's present state of mind—that could impair effective use even of one's own judgments. Thus, investigating how decision-makers use multiple opportunities to estimate the same quantity reveals not only whether and how effectively individuals can apply the normatively correct strategy of combining those estimates, it can also indicate the broader mechanisms by which people make use of multiple, potentially conflicting judgments.

In the present study, we assessed how—and how effectively—decision-makers use several judgments made in response to the same world knowledge question. In particular, we contrast two bases on which participants might decide how to choose or combine those judgments: (a) the plausibility of particular individual estimates and (b) general naïve theories about the value of averaging and of early and later judgments (Soll, 1999). We ask whether metacognition about multiple estimates is more effective given cues supporting one basis or the other—or both together—and what differential performance across cues reveals about the metacognitive bases for such decisions.

The wisdom of crowds and the crowd within

Individuals are frequently called upon to make quantitative estimates, such as projecting a business's sales, forecasting the temperature, judging the time needed to complete a project, or simply answering general knowledge questions such as *What percent of the world's population is 14 years of age or younger?* These estimations have been modeled (Yaniv, 2004) as a function of three sources: (a) the true value, (b) a systematic bias on the part of the judge to respond too high or too low, and (c) random error, such as variability in how knowledge is retrieved or translated into an estimate.

As long as random errors are at least partially independent, averaging multiple estimates reduces the influence of those errors (Yaniv, 2004). In addition, when bias varies across judges, averaging also reduces this bias towards the mean bias present in the population; this also improves accuracy unless some judges are substantially less biased than the rest of the population and can be identified as such (Soll & Larrick, 2009). Consequently, the average of multiple judges is at least as accurate as the average judge and can often outperform *any* judge, especially[1] in cases where the judges *bracket* the true value, or provide estimates on either side of the answer (Soll & Larrick, 2009). For example, suppose that one judge estimated that 40% of the world's population was under 14 years of age and a second judge estimated that only 20% was. In this case, averaging the judges' responses produces an estimate of 30%, which is closer to the true value of 26% (Central Intelligence Agency, 2011) than either original judge. This phenomenon has been demonstrated in a long-standing literature showing that quantitative estimates can be made dramatically more accurate by aggregating across multiple judges (Galton, 1907), a principle often termed *the wisdom of crowds* (Surowiecki, 2004).

The same principles apply even to multiple estimations from the same individual. Although individuals may be consistent in their bias, any stochasticity in how individuals sample their knowledge or translate it into a numerical estimate still produces random error, and this error can be reduced by averaging over multiple estimates.[2] Thus, the average of multiple estimates even from the same individual typically outperforms any of the original estimates (Vul & Pashler, 2008). This difference has been termed the benefit of the *crowd within* (Vul & Pashler, 2008) and has been argued to support a view in which judgments are based on probabilistic rather than deterministic access to knowledge (Vul & Pashler, 2008; see also Hourihan & Benjamin, 2010; Koriat, 1993, 2012; Mozer, Pashler, & Homaei, 2008).

Because multiple estimates from the *same* individual are less independent (that is, are more strongly correlated) than estimates from different individuals, averaging within an individual does not decrease error as much as averaging between individuals (Müller-Trede, 2011; Rauhut & Lorenz, 2010; Vul & Pashler, 2008). Nevertheless, as long as the estimates are even *partially* independent of one another, the technique still confers a benefit (Vul & Pashler, 2008). Furthermore, the benefits increase when the two guesses are less dependent on one another—as is the case when the second judgment is delayed (Vul & Pashler, 2008; Welsh, Lee, & Begg, 2009), when individuals' low memory span prevents them from sampling as much of their knowledge at one time (Hourihan & Benjamin, 2010), or when participants[3] are encouraged to re-consider assumptions that might have been wrong (*dialectical bootstrapping*; Herzog & Hertwig, 2009; for further discussion, see Herzog & Hertwig, 2013; White & Antonakis, 2013).

Knowing the crowd within

Despite the substantial benefits of aggregating multiple estimates, decision-makers consistently undervalue this strategy when it comes to averaging across multiple judges. When asked to reason explicitly about the values of averaging, they often assume that the average performs no better than the average judge (Larrick & Soll, 2006); in reality, as reviewed above, the average often outperforms *any* judge. And, when allowed to make judgments informed by one or more other individuals' estimates, participants tend to inappropriately discount the advice of others rather than productively combining the advisor's knowledge with their own (for review, Bonaccio & Dalal, 2006).

In particular, decision-makers appear to rely on a *choosing* strategy (Gigerenzer & Goldstein, 1996) of using only a single cue—often one's own estimate—rather than attempting to combine multiple cues, such as estimates made by several different judges (Soll & Larrick, 2009). Choosing can be effective when the best cue or judge can be easily identified and when the estimates are not particularly independent (i.e., are strongly correlated), so that there is little random error to reduce through averaging (Soll & Larrick, 2009). However, individuals are often ineffective at actually determining the best judge (Soll & Larrick, 2009), and in situations that involve estimates from different individuals, the estimates are often sufficiently independent that averaging outperforms even choosing the best judge with perfect accuracy (Soll & Larrick, 2009). It has thus generally been concluded that decision-makers underuse a strategy of averaging several individuals' estimates even in environments where it would be helpful (Bonaccio & Dalal, 2006; Harvey &

[1] The exact relation of the average of the estimates to the average judge depends on how accuracy and inaccuracy are quantified (Soll & Larrick, 2009). If inaccuracy is quantified as the absolute deviation from the true value, the average outperforms the average judge only when the judges bracket the true value; such instances can be quite frequent when averaging between individuals (Soll & Larrick, 2009). If inaccuracy is quantified as squared error, averaging can outperform the average judge even without bracketing because squared error particularly penalizes large deviations from the true value, and averaging reduces the influence of these extreme estimates. We focus here on squared error to facilitate comparison with past examinations of within-person averaging (e.g., Herzog & Hertwig, 2009; Vul & Pashler, 2008), which have used squared error, but all of the qualitative results hold when absolute deviation is considered instead.

[2] This principle holds so long as the samples are drawn from the same internal distribution. If the mean or variance of this distribution shifts over time naturally or as a consequence of the decision task, aggregating estimates could result in less accurate estimations (Rauhut & Lorenz, 2010).

[3] In principle, it is possible that participants might prefer to aggregate their estimates in some other way, such as a weighted average in which, for instance, the first estimate receives a weight of two-thirds and the second estimate a weight of one-third. We included only the unweighted average as a response option for three reasons. First, assigning equal weight to each cue has been previously proposed as a normative strategy (*unit weighting*; Einhorn & Hogarth, 1975). Second, most of the benefits of combining cues accrues from using the cues at all, with the exact weights assigned to the cues contributing relatively little (the *flat maximum effect*; Lovie & Lovie, 1986). Finally, even when participants are allowed to freely choose weights for their own and others' estimates, they rarely assign unequal weightings (Soll & Larrick, 2009).

Fischer, 1997; Mannes, 2009; Soll & Larrick, 2009; Yaniv, 2004; Yaniv & Choshen-Hillel, 2012).

Why do decision-makers underuse a strategy as simple and powerful as averaging the estimates of multiple judges? Some explanations have focused on the social aspects of working with multiple judges, such as a belief that one is better than the average judge (Harvey & Fischer, 1997; Lim & O'Connor, 1995) or the fact that individuals know the reasons for their own judgments but not those of others (Yaniv, 2004). These biases are less applicable to within-person averaging, and such accounts predict that participants may combine their own judgments even though they undervalue between-person combinations. However, other explanations of the tendency against between-person averaging predict a similar aversion to within-person averaging. For instance, one proposal is that many people hold incorrect naïve theories about the statistical benefits of averaging (Soll, 1999); such theories would discourage both types of averaging. Both types of averaging might also be influenced by the temporal ordering of the judgments (Hogarth & Einhorn, 1992): in both types of averaging, individuals are presented with an estimate more distant from their present state of mind—either their own estimate at an earlier point in time or another judge's estimate—and an estimate that is closer to it.

Thus, whether or not individuals are similarly reluctant to average their own estimates can inform more general theories of how decision-makers reason about multiple, possibly conflicting judgments. Moreover, the willingness of decision-makers to average their estimates also has direct applied value because there is interest in improving the accuracy of judgments through multiple estimations (Herzog & Hertwig, 2009) or related methods (such as more-or-less estimation; Welsh et al., 2009).

Some evidence suggests that decision-makers may indeed underuse within-person averaging. Müller-Trede (2011) asked participants to make a third estimate while viewing their first two estimates and found that, as with between-person averaging, participants often retained one of the original estimates rather than aggregating them. However, it is not yet clear how participants made this decision or what caused their dispreference for averaging. In the present study, we investigate the metacognitive basis of decisions about combining multiple self-generated estimates and how those may or may not parallel the bases underlying decisions from multiple individuals.

Making metacognitive judgments

The evidence suggests that metacognitive decisions can be made on multiple bases, some of which are more effective for a particular judgment than others. In particular, theories of metacognition (e.g., Kelley & Jacoby, 1996; Koriat, 1997) have often distinguished judgments made on the basis of general naïve theories from judgments made on the basis of the subjective experience of interacting with a particular item. This distinction is supported by dissociations in metacognition between participants' general beliefs and their judgments about specific items. For example, participants state a general belief that memory for words will decrease over time, but their predictions of their ability to remember individual words within an experiment at a particular point in the future are not always influenced by the time that will elapse before the test (Koriat, Bjork, Sheffer, & Bar, 2004; but see Rawson, Dunlosky, & McDonald, 2002). But, participants directly compare time points, their predictions are more apt to accurately incorporate forgetting (Koriat et al., 2004). Similarly, although people state that studying words multiple times will improve their memory, their predictions of their ability to remember a specific item are not very sensitive to how many times that item will be studied (Kornell & Bjork, 2009; Kornell, Rhodes, Castel, & Tauber, 2011).

Whether a judgment is made based on item-specific properties or based on a general belief may depend on the cues in the decision environment. For example, Kelley and Jacoby (1996) asked participants to rate how difficult it would be to solve particular anagrams (e.g., unscrambling *fscar* to form *scarf*). When participants had to first solve the anagrams on their own, they could use their own feeling of ease or difficulty in solving the item to judge its difficulty. Ratings made on this basis were fairly predictive of how successfully others could solve each anagram. However, when the task displayed the correct answer from the start, they could no longer rely on their own experience solving that particular item, and had to turn to other bases for judgment, such as general beliefs about what factors make anagrams difficult. These ratings less accurately predicted how well others could unscramble the anagrams.

Although the anagrams are a situation in which item-based responding produces better estimates than a naïve theory, the reverse is often true: One's experience with a particular item is sometimes influenced by factors inversely rated or unrelated to the property being judged, which can introduce systematic bias into the decision process (Benjamin & Bjork, 1996). For example, Benjamin, Bjork, and Schwartz (1998b) asked participants to learn short lists of word pairs and judge their future ability to recall each pair. The last pair in a list, which was most recent and active in memory at the time of the judgment, was judged to be the most memorable. However, over the long term, the benefits of recency fade in favor of a benefit for items studied first (the recency-to-primacy shift; Postman & Phillips, 1965), so that the recent pairs, which participants judged as most memorable, were actually *least* apt to be remembered later. That is, judgments of whether items were memorable were systematically inaccurate in this task because the judges' experience with each item was influenced by properties inversely related to the outcome they were attempting to predict.

However, as will become relevant later, misinterpretations of item-level experience can be restrained when the feeling of fluency can be correctly attributed to its true source. For example, imposing a heavy perceptual mask makes words harder to read and thus less apt to be judged as previously studied in a recognition memory task. But if participants are warned about the effect beforehand, they can correctly attribute the lack of fluency to the perceptual mask, and its influence on memory judgments disappears (Whittlesea, Jacoby, & Girard, 1990).

Decisions about how to use multiple estimates could plausibly be made on either the basis of a general theory or on item-specific judgments, and it is not clear *a priori* which would be more effective. For instance, participants might aggregate their estimates on the basis of having an accurate naïve theory about the value of such a strategy. However, theory-based responding could also produce poor judgments if participants held an inaccurate naïve theory: much of the benefit of within-person averaging derives from reducing random error, but many individuals do not appreciate that averaging helps cancel out random sources of error (Larrick & Soll, 2006; Soll, 1999) and so may not have reason to combine their estimates. Similarly, responding based on the characteristics of a particular estimate could be effective if participants can use item-level knowledge to identify the most accurate estimate, but it could also be misleading if item-level factors such as fluency or mnemonic accessibility biased participants towards a particular estimate—for instance, the one made most recently—whether it was right or wrong.

Present study

In four studies, we examined how—and how effectively—participants decide how to use multiple estimates. We assessed whether participants exhibited a similar underuse of within-person averaging as they do between-person averaging, and, to investigate the source of any such bias, we tested whether the effectiveness of these metacognitive decisions varied as a function of whether they were made on the basis of general beliefs, item-specific evaluations, or both.

Following Vul and Pashler (2008), we asked participants to estimate answers to general knowledge questions, such as *What percent of the world's population is 14 years of age or younger?*, and then later unexpectedly asked them to make a second, different estimate. As will be seen, the average of these two estimates tended to be more accurate than either estimate by itself, replicating prior results (Rauhut & Lorenz, 2010; Vul & Pashler, 2008). In a new third phase, we then asked participants to select their final response from among their first guess, second guess, or average.

The information present during this third phase varied across studies to emphasize different bases for judgment. In Study 1, we randomly assigned participants to one of two conditions. One condition provided cues intended to emphasize participants' general beliefs about how to use multiple estimates, and the other condition provided cues emphasizing item-specific evaluations. For ease of exposition, we present these conditions as Study 1A and Study 1B, respectively, before comparing the results across conditions. Next, in Study 2, we further tested hypotheses about participants' use of cues emphasizing item-specific evaluations. Finally, Study 3 provided both theory-based and item-specific cues together in the third phase.

In each study, we examined the consequences of these cues on two aspects of participants' decision-making. First, we examined the *decisions* made by participants: did they employ an averaging strategy, or did they choose one of their original responses? Second, we tested whether participants made these strategy decisions effectively by examining the *accuracy* of the answers they selected. We calculated the mean square error (MSE) of participants' final answers by computing, for each trial, the squared deviation between the true answer to the question and the particular estimate selected by the participant. We then compared this MSE to the MSE that would have been obtained under several other strategies, such as always averaging or selecting randomly among the three available options.

This analytic strategy allowed us to examine the effectiveness of participants' selections at two levels. First, participants might (or might not) exhibit an overall preference for the strategy that yields the best performance; based on prior results (Rauhut & Lorenz, 2010; Vul & Pashler, 2008), we predicted this overall best strategy to be averaging. However, the average may not be the optimal selection on *every* trial. When estimates are highly correlated, as is the case for within-individual sampling (Vul & Pashler, 2008), averaging can be outperformed on some trials by choosing one of the original estimates (Soll & Larrick, 2009). Thus, a second level at which performance can be analyzed is whether participants adopt particular strategies (such as averaging) selectively on those trials for which those strategies would be most accurate (as has been observed in other tasks; e.g., Payne, Bettman, & Johnson, 1988). We term the adoption of particular strategies for particular trials *trial-by-trial strategy selection.*

Study 1

In Study 1, we varied the cues provided to participants when they decided whether to choose or combine estimates. After making a first estimate for each item and then a second estimate, all participants decided, separately for each item, whether to submit their first guess, their second guess, or the average of their two guesses. However, the way these three final response options were presented was manipulated between participants.

Participants randomly assigned to the labels-only condition (Study 1A) saw the three response options described with the labels *your first guess*, *your second guess*, or *the average of your two guesses* on all trials; participants did not see the particular numerical values represented by the first guess, second guess, and average. This decision environment would be expected to encourage participants to apply their general beliefs about averaging versus choosing strategies, but provides little opportunity to evaluate the fluency or subjective plausibility of particular estimates at the item level.

By contrast, participants in the numbers-only condition (Study 1B) saw only the specific numerical values that they had previously provided and never received any information that these three values represented their first estimate, second estimate, and average estimate. Because the numbers-only task does not include explicit descriptions of when or how the numerical estimates were obtained, we expected that participants would be likely to rely less on their naive theories about the effects on those variables on accuracy. Instead, participants would have an item-level basis for responding: the subjective plausibility or fluency of each number as an answer to the question. Potentially, this item-specific information could support more accurate metacognition if the true answer seemed particularly plausible to participants (e.g., because it should be closer to the mean of the distribution of their samples of knowledge). Because the particular numeric estimates vary from trial to trial (unlike the labels), they might also provide a basis for trial-by-trial strategy selection. Alternately, these item-based judgments might be *less* effective than the theory-based judgments in Study 1A if participants' item-level perceptions are contaminated by misleading sources of fluency, such as the recency or subjective plausibility of the original estimates.

Method

Participants

In this and all subsequent studies, participants were students at the University of Illinois or members of the surrounding community who participated for course credit or a cash honorarium. One hundred and twelve people participated in Study 1; sixty-one were randomly assigned to the labels-only condition (Study 1A) and fifty-one of the Study 1 participants were randomly assigned to the numbers-only condition (Study 1B) condition.

Materials

Twelve questions assessed participant's knowledge of worldwide demographic characteristics or of statistics regarding particular countries, such as *What percent of the world's population is 14 years of age or younger?* Eight of the items were those used by Vul and Pashler (2008), and four similar items were created from *The world factbook* (Central Intelligence Agency, 2011). The items were selected so that participants could make an informed estimate from world knowledge but were unlikely to be able to retrieve an exact answer from memory. The answers to all of the questions were percentages and thus on the same scale. We used the same materials in all studies reported here. Table 1 reports the items and their answers.

Procedure

The first two phases of the experiment were the same across conditions. In the initial phase, all participants were presented with the questions one at a time on the computer and typed their best guess to answer each question. To encourage participants to read and think about each question, after the presentation of the question, a 1500 ms delay was enforced before participants could begin to type a response.

Importantly, no information was presented during the first phase that indicated participants would ever make a second estimate for any of the items. Rather, participants were instructed to *provide your best guess of the correct answer for each fact provided*. These manipulations make it unlikely that participants would use their two response opportunities to provide the endpoints of a range (Vul & Pashler, 2008).

After making a first estimate for all 12 questions, participants completed an unrelated episodic memory task for approximately 30 min. Participants were then told that, because some of the general knowledge questions from earlier were rather difficult, they were being given another opportunity to answer the questions. Participants were explicitly instructed to *type a second, different guess for the question*; following Vul and Pashler (2008), these instructions were chosen so that participants did not think the goal of the task was to match their second estimate to their first. Participants answered each question a second time in a re-randomized order using the same procedure as the first phase.

After a second estimate had been made for all 12 questions, the participants immediately proceeded to the third and final phase. For this final decision phase, participants saw each question a third time, with the order of presentation again re-randomized. For each item, participants clicked on one of three boxes with the mouse to indicate what they wanted to report as their final decision. The instructions and available cues were manipulated across conditions. In the labels-only condition (study 1A), participants were instructed that, *Now that you have made two different guesses on these questions, it is time to choose what to submit as your final answer.* Participants were told that they could submit their first guess, average guess, or second guess, and an example of these calculations was provided. For each item, three boxes labeled *Your first guess*, *The average of your two guesses*, and *Your second guess* were displayed. Only these labels were displayed; the particular numerical values that these labels represented for each trial were not shown

By contrast, in the numbers-only condition (study 1B), participants were told only that they would have *a multiple-choice decision between three possible answers*. Then, on each trial, rather than the labels *first guess, average*, and *second guess*, the three options participants could select among were the numerical values (rounded to the nearest integer) of the first estimate, average, and second estimate. No mention was made at any point that the values came from the participants' prior guesses or the average thereof.

To control for any effects of how the response options were ordered on the screen, the same spatial order was used in both conditions: the first estimate, then the average, and then the second estimate. As in the previous phase, a 1500 ms delay was enforced between the presentation of the stimulus question and the appearance of the response boxes.

In some trials of both studies 1A and 1B, participants provided estimates that differed by fewer than two percentage points. In these cases, the first, second, and average estimate did not constitute three distinct integer values. (For example, averaging original estimates of 50 and 49 produces 49.5, which is not distinct from the two original estimates when rounded to an integer.) Because participants rarely provided estimates at greater than integer precision (fewer than 1% of trials), these trials would include in the final decision phase values that were essentially identical from the participant's perspective. To ensure that any potential benefits of averaging were not driven purely by whether participants made two effectively identical estimates, trials in which the initial estimates did not differ by at least two percentage points were discarded and not re-presented to participants during the third phase (for further discussion, see Herzog & Hertwig, 2013; White & Antonakis, 2013).

Results

We report three aspects of participants' judgment and decision-making. First, we present participants' performance in the initial judgment tasks; these tasks did not differ across conditions. Next, we characterize participants' metacognitive performance in the final decision phase in each of the two conditions (numbers-only and labels-only). Finally, we present a direct comparison of participants' performance given one cue type versus the other. Each analysis afforded comparisons to multiple potential baselines; in the text, we focus on those comparisons that were relevant to the hypotheses of interest in each study, but we use the tables and figures to provide a full characterization of participants' behavior in each phase of the task.

Accuracy of estimates

Table 2 presents the accuracy of participants' estimates in this and the other present studies. Overall, participant's first estimates (*MSE* = 531, *SD* = 349) had lower squared error (that is, were more accurate) than their second (*MSE* = 619, *SD* = 380), $t_{(111)} = -3.21$, $p < .01$, 95% confidence interval of the difference: [−141, −33]. But, the average of the two estimates (*M* = 501, *SD* = 320) was more accurate still and outperformed even the first estimate, $t_{(111)} = -2.05$, $p < .05$, 95% confidence interval of the difference: [−60, −1].

Importantly, the fact that the second estimate enhanced accuracy when combined with the first indicated that it contributed new, previously unused information. If the second estimate had been pure noise (i.e., participants typed in a purely arbitrary value when required to make a second estimate), it would not have been useful to combine with the first. Thus, this result replicates the benefit of combining multiple estimates from the same judge (Vul & Pashler, 2008). The crucial question in the present study, however, was whether participants would recognize this benefit and select the average as their final answer.

Study 1A (labels only)

Participants in Study 1A saw only the labels in the final decision phase. 27% of trials in Study 1A were omitted from the third phase because the estimates differed by fewer than two percentage points for reasons described above.

Final selections. Participants' selections in the final reporting phase of each study are depicted in Table 3. Overall, participants in Study 1A reported the average most frequently (*M* = 59% of trials, *SD* = 28%), more than they chose their first guess (*M* = 19%, *SD* = 19%) or chose their second guess (*M* = 22%, *SD* = 23%). A one-sample *t*-test revealed the rate of averaging was reliably greater than the 33% that would be expected from chance selections, $t_{(60)} = 7.30$, $p < .001$, 95% confidence interval of the mean: [52%, 66%].

However, few participants exclusively adopted either an averaging strategy or a choosing strategy. Fig. 1 displays a histogram of the proportion of times each participant selected the average and reveals that the majority of subjects applied averaging to some trials and a choosing strategy to others. This raises the possibility that participants may have effectively modulated their strategy on a trial-by-trial basis, adopting an averaging versus choosing strategy depending on what would be most effective for a particular decision environment. We test this hypothesis below.

Performance of strategies. To assess the effectiveness of participants' decision strategies, we computed the mean squared error (MSE) of the final response selected on each trial (that is, whichever of guess 1, guess 2, or the average was selected). We compared this value to the MSE that each participant would have obtained by applying several alternate decision strategies to those same trials. *Ideal decision-making* is the MSE that would result if a participant selected with perfect accuracy, on a per-trial basis, whichever of the three response options had the lowest error. The ideal decision-making value defines the upper bound of performance in the metacognitive task, analogous to an ideal observer (e.g., Peterson, Birdsall, & Fox, 1954) in a psychophysical task. Note that even perfect metacognition would not result in an MSE of 0 because even the best of the three options rarely corresponded to the exact answer to the world knowledge question. *Random responding* was the expected value of selecting randomly with equal probability among the three options. This value provides a baseline that would be obtained if participants had no metacognitive insight. However, participants could actually underperform even this baseline if they had an ineffective metacognitive strategy that led them to systematically select suboptimal estimates.

Three other values were calculated to characterize the averaging and choosing strategies. *Always average* was the MSE that would be obtained by averaging on every single trial. *Random choosing* was the expected value of always applying a choosing strategy but choosing randomly between the two original estimates; that is, it was average squared error of the two guesses on each trial. *Perfect choosing* was the MSE of always applying a choosing strategy and always choosing the better original estimate (but never averaging). Thus, it was the MSE of the more accurate of the participants' two original estimates on each trial.

Finally, what we term the *proportional random* strategy was the expected value of each participant selecting the same proportion of the three response types (first guess, second guess, and average) as they actually selected, but with those proportions randomly assigned to the twelve trials. For example, for a participant who selected the first estimate 20% of the time, the second estimate 30% of the time, and the average 50% of the time, the proportional random strategy would be the expected value of selecting the first guess on a random 20% of trials, the second guess on a random 30% of trials, and the average on a random 50% of trials. The proportional random strategy would be equivalent to the participant's observed performance if and only if participants had assigned their mix of strategy choices arbitrarily to particular trials; e.g., in a probability matching (Friedman et al., 1964) strategy. However, if participants effectively selected strategies on a trial-by-trial basis—for example, by being more apt to average on trials for which averaging was indeed the best strategy—then participants' actual selections would outperform the proportional random strategy.

Table 2
Mean squared error of participants' first, second, and average estimate in each of the present studies.

	Estimate 1		Estimate 2		Average	
	M	*SD*	*M*	*SD*	*M*	*SD*
Study 1	531	349	619	380	501	320
Study 1A	565	377	658	416	532	368
Study 1B	491	310	572	330	464	250
Study 2[a]	588	372	649	428	560	349
Study 3						
Average-middle display	553	363	578	323	485	292
Average-last display	458	325	511	370	424	315
Mean of two display types	504	344	543	346	453	303
All studies combined	537	338	606	376	500	312

Note: SD = standard deviation.

[a] Although participants in Study 2 made these estimates in the initial phases, they did not see them in the final decision phase; instead, they decided among the estimates of a Study 1B participant to whom they were yoked.

Table 3
Rates at which participants selected the first estimate, second estimate, or average estimate as their final response in each of the present studies, and accuracy on choosing the better estimate on trials in which the participants selected one of the original two estimates.

	Estimate 1		Estimate 2		Average		Choosing accuracy[a]	
	M (%)	*SD*	*M* (%)	*SD*	*M* (%)	*SD*	*M* (%)	*SD*
Study 1A (labels)	19	19	22	23	59	28	57	28
Study 1B (numbers)	23	15	34	19	43	17	47	23
Study 2 (yoked numbers)	36	19	28	16	36	20	57	27
Study 3 (labels and numbers)								
Average-middle display	22	25	31	19	47	21	61	26
Average-last display	25	20	31	26	44	23	51	16
Mean of two display types	24	23	31	23	45	22	56	23

Note: SD = standard deviation.

[a] Calculated for each participant based on trials in which the participant selected one of the original estimates (rather than the average) and in which one estimate was closer to the true answer than the other.

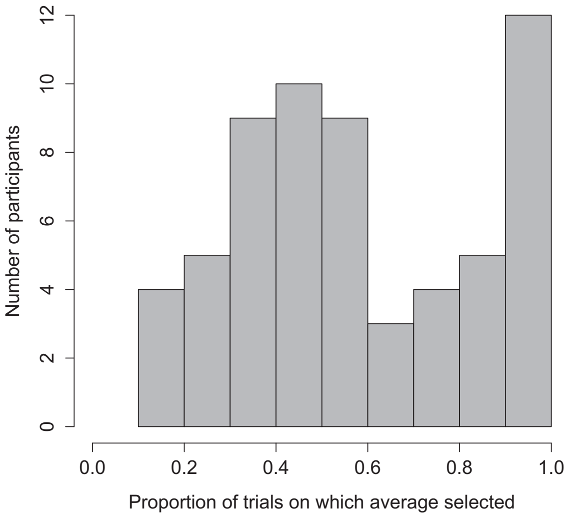

Fig. 1. Histogram of the proportions of trials on which participants averaged their two estimates in Study 1A.

The squared error that would be obtained in Study 1A under each of these strategies, as well as participants' actual accuracy, is plotted in Fig. 2.

Given just the strategy labels, participants' actual selections ($MSE = 561$, $SD = 374$) outperformed randomly selecting among all three options ($MSE = 584$, $SD = 371$), $t_{(60)} = -2.17$, $p < .05$, 95% CI of the difference: [−45, −2]. This result indicates that participants had some metacognitive awareness that enabled them to select among options more accurately than chance.

However, participants' responses resulted in greater error than a simple strategy of always averaging ($MSE = 541$, $SD = 368$), $t_{(60)} = 2.53$, $p < .05$, 95% CI: [6, 53]. Participants performed even worse relative to perfect choosing between the two original estimates ($MSE = 373$, $SD = 296$), $t_{(60)} = 10.28$, $p < .001$, 95% CI: [157, 232]. (Averaging outperforms perfect choosing of the better original estimate only when the estimates bracket the true answer with sufficient frequency,[4] but the bracketing rate was fairly low at 26%.)

[4] Estimates made by *different* individuals can bracket the true value at rates of 40% or higher (e.g., Soll & Larrick, 2009); in such situations, averaging can outperform even perfect choosing. The lower rate of bracketing when averaging multiple within-person estimates is expected because estimates from the same individual are more correlated with each other than estimates from different individuals and are thus less likely to bracket the true value. As will be seen later, however, even when averaging does not outperform perfect choosing, averaging can be an effective strategy because it does not require individuals to be able to actually identify their better guess.

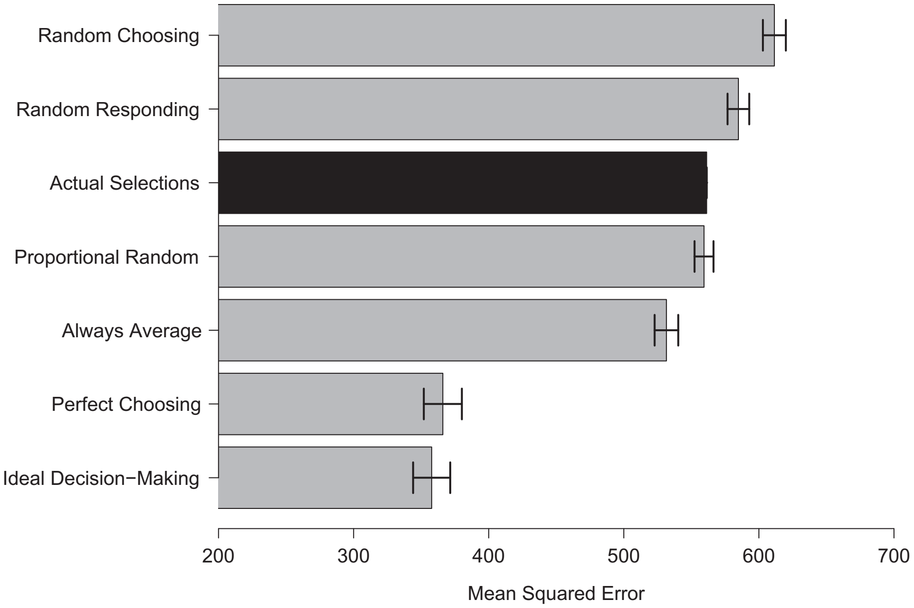

Fig. 2. Mean squared error (MSE) of participants' final selections in Study 1A versus the MSE that would have been obtained under several comparison decision strategies. Error bars indicate the 95% confidence interval of the difference in MSE between participants' actual selections and each alternate strategy.

Moreover, there was no evidence that participants were effectively selecting strategies on a trial-by-trial basis. Participants' responses did not result in lower squared error than the proportional random strategy (*MSE* = 568, *SD* = 372), $t_{(60)} = 0.20$, $p = .84$, 95% CI: [−17,21]. This cannot be attributed simply to insufficient statistical power because participants' selections actually resulted in numerically *higher* squared error than the proportional random baseline.

Interim discussion. Study 1 assessed participants' metacognition about how to use multiple self-generated estimations by asking participants to decide, separately for each question, whether to report their first estimate, their second estimate, or the average of their estimates. In Study 1A, participants made this selection under circumstances that emphasized their general beliefs about the merits of these strategies: Participants viewed descriptions of the response strategies but not the particular numerical values that those options represented for each item.

Combining estimates was useful, and participants recognized this to some degree. Replicating previous results, the average of the two estimations was somewhat more accurate than either of the estimates themselves. Participants showed some evidence for metacognitive appreciation of this benefit in that they selected the average as their final response more than the other options and consequently outperformed a random selection among the options.

But Study 1A also revealed limits to participants' metacognition. Although participants did show some preference for the average, they could have produced more accurate reporting had they averaged even more frequently. Moreover, although it is possible to imagine that participants could have had a naïve theory that led them to average on some trials and choose on others (e.g., if they had a theory that certain types of questions would benefit from averaging more than others), they did not actually show any ability of effective trial-by-trial strategy selection. They performed no better than selecting the same proportion of strategies on a random set of trials.

Thus, the results of Study 1A suggest that in a decision environment emphasizing participants' general beliefs about how to use multiple judgments, participants have some preference for combining those judgments, albeit a weak one, but no apparent ability to select strategies on a trial-by-trial basis. In Study 1B, we contrast this with participants' decisions in an environment emphasizing item-level decisions.

Study 1B (numbers only)

In the final decision phase of Study 1B, participants saw only the numerical values represented by the first estimate, second estimate, and average. As in Study 1A, trials in which participants' initial estimates[5] differed by less than two percentage points (24% of trials) were excluded from the final decision phase because the first estimate, average, and second estimate did not constitute three distinct integer values to decide among.

Final selections. Participants showed a somewhat different pattern of selections in the third phase when only the

[5] There was some variability across studies in how close participants' initial estimates were to the true value. These differences were presumably spurious because the studies were identical up until the final decision phase.

numerical cues were provided. As in Study 1A, participants selected the average ($M = 43\%$) more than the first guess ($M = 23\%$) or second guess ($M = 34\%$). This rate of averaging was greater than would be expected by chance, $t_{(50)} = 4.06$, $p < .001$, 95% CI of the rate: [38%,48%], but it was lower than in Study 1.

To further characterize participants' selections, we examined the trials on which participants chose one of the original estimates rather than average. They were no better than chance at identifying the better of the two estimates. It was not that participants merely improved over chance by a degree too small to be statistically reliable. Rather, they were actually numerically more apt to choose the *worse* of the two estimates: the more accurate estimate was selected on only 47% of choosing trials (95% CI: [40%,53%]) and the less accurate on 53%, $t_{(50)} = -.99$, $p = .33$.

Performance of strategies. Fig. 3 plots the squared error of participants' actual final selections and the comparisons to the alternate strategies described above.

The differing pattern of selections in Study 1B had consequences for the accuracy of participants' reporting. In Study 1B, participants' actual selections ($MSE = 517$, $SD = 294$) did not show less error than responding completely randomly ($MSE = 508$, $SD = 267$). In fact, participants' responses had a numerically *greater* squared error than even purely random responding although this difference was not statistically reliable, $t_{(50)} = 0.59$, $p = .56$, 95% CI; [−20,37].

Comparison of cues

The results presented above reveal that participants who saw the strategy labels (Study 1A) reliably outperformed random selection, but that participants who saw numerical estimates (Study 1B) did not. As noted previously, participants in Study 1 were randomly assigned to see one cue type or the other. This allowed us to test the effect of this between-participant manipulation of cues by directly comparing participants' metacognitive performance between conditions.

Note that the previously presented comparisons between participants' actual strategies and the comparison strategies were within-participant comparisons that inherently controlled for the overall accuracy (MSE) of each participant's original estimates. However, a *between*-participant comparison of the raw MSE of participants' final selections could also be influenced by individual differences in the MSE of the original estimates that participants were deciding among. Indeed, participants varied substantially in the accuracy of their original answers to the world knowledge questions. As our primary interest was in participants' metacognitive decisions about the estimates in the final reporting phase and not in the general accuracy of the original estimates, a desirable measure would control for such differences in baseline accuracy. By analogy to Mannes (2009) and Müller-Trede (2011), we computed a measure of how effectively each participant, given their original estimates, made use of the opportunity to select among the first estimate, second estimate, and average. We calculated the percentage by which participants' selections overperformed (or underperformed) random selection; that is, the difference in MSE between each participant's actual selections and random selection, normalized by the MSE of random selection.

A comparison across conditions of participants' gain over random selection confirmed that the labels resulted in better metacognitive performance than the numbers. Although participants in the labels-only condition (Study 1A) improved over random selection ($M = 5\%$ reduction in MSE), participants in the numbers-only condition (Study 1B) underperformed it ($M = -2\%$). This difference was reliable, $t_{(110)} = 1.99$, $p < .05$, 95% CI of the difference: [15%,1%].

Why was participants' metacognition less effective in Study 1B than in Study 1A? We conducted two comparisons of the final response options chosen by participants. First, participants were reliably less likely to average in Study 1B (43% of trials) than in Study 1A (59%), $t_{(110)} = -3.60$, $p < .001$, 95% CI of the difference: [−25%,−7%]. Given that participants could have obtained substantially lower error by simply averaging on all trials, the reduced rate of averaging in Study 1B contributed to the increased error of participants' reporting. Second, there was also some evidence that the Study 1B participants were also less successful at implementing the choosing strategy. When participants chose one of the original estimates rather than average, they were more successful at choosing the better of the two estimates in Study 1A (57% of choosing trials) than in Study 1B (47% of choosing trials); this difference was marginally significant, $t(98) = -1.91$, $p = .06$, 95% CI of the difference: [−20%,0%].

Discussion

In Study 1B, we assessed participants' metacognition about how to choose or combine multiple estimates when presented with a decision environment emphasizing item-based decisions. Participants saw the numerical values represented by their first estimate of a world fact, their second estimate, and the average of these two estimates, but no explicit labels of these strategies. This decision environment resulted in reliably less effective metacognition than the cues in Study 1A, which emphasized theory-based decisions. First, participants were less apt to average their estimates in Study 1B than in Study 1A; this reduced the accuracy of their reports because averaging was typically the most effective strategy. There was also some evidence that, when participants chose one of the original estimates rather than average, they were less successful at choosing the better estimate in Study 1B than in Study 1A. In fact, the Study 1B participants were numerically less accurate than chance at choosing the better estimate. Consequently, unlike in Study 1A, the accuracy of participants' final estimates was not reliably better than what could have been obtained from purely random responding. A simple strategy of always averaging could have resulted in substantially more accurate decisions.

The differing results across conditions provide evidence against two alternate explanations of the results thus far. Because the order of the response options was fixed, a less interesting account is that participants' apparent preference for the average in Study 1A, or their preference for

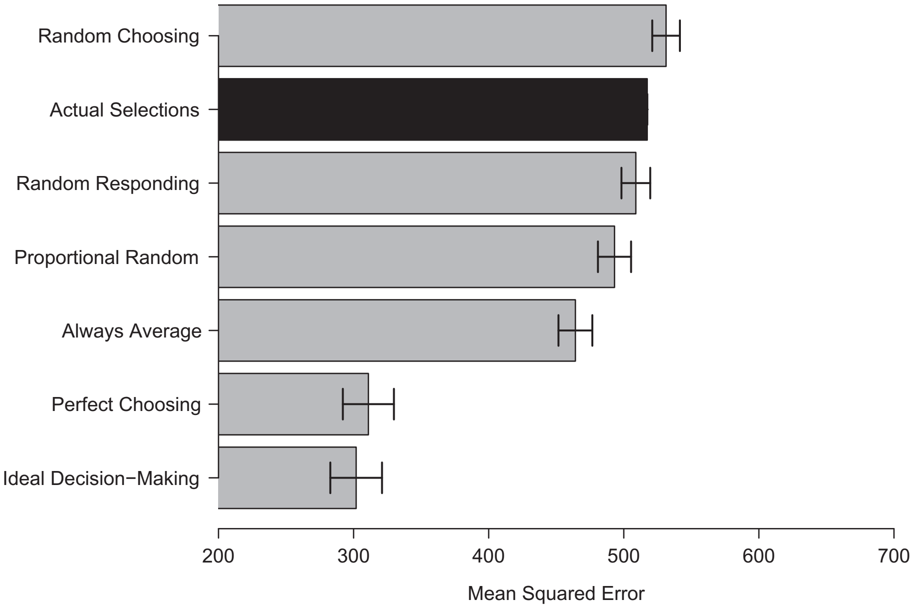

Fig. 3. Mean squared error (MSE) of participants' final selections in Study 1B versus the MSE that would have been obtained under several comparison decision strategies. Error bars indicate the 95% confidence interval of the difference in MSE between participants' actual selections and each alternate strategy.

their second guess in Study 1B, was driven purely by the locations of those options on the screen. However, this account cannot explain why participants' degree of preference for each option, and the accuracy of their decisions, differed across studies given that the response options were located in the same position in both studies. (Study 3 will provide further evidence against this hypothesis by experimentally manipulating the location of the choices in the display.) Second, it is possible in principle that participants given the labels in Study 1A did not decide primarily on the basis of a general naïve theory about the benefits of averaging versus choosing, but rather on an item-level basis. Participants could have retrieved or calculated the numerical values associated with each of the labels *first guess*, *second guess*, and *average guess* and then assessed the plausibility of those values. Conversely, participants in Study 1B could have identified the three numerical values as their first, second, and average estimate and responded on the basis of a naïve theory about those strategies. The divergence in metacognitive performance across studies, however, indicates that participants did not approach the task identically across studies; presenting different information at the time of the final decision altered participants' decisions and accuracy.

The contrast between Studies 1A and 1B, then, provides evidence that metacognitive decisions about using multiple estimates can be made on different bases and that these bases vary in their effectiveness. When participants saw descriptions of the strategies in Study 1A, they could easily apply their naïve theories about the effectiveness of those strategies. This environment was somewhat effective at promoting an averaging strategy and thus allowing participants to make accurate reports. However, when participants were given only three numerical estimates to select among, there was little information available that could support a decision based on those theories. Rather, participants likely had to rely (or rely to a greater degree) on assessments of the numbers on individual trials, perhaps on the basis of the numbers' fluency or subjective plausibility. Under these circumstances, participants were less apt to select the average, and the estimates they reported as their final selections were no more accurate than what would be obtained from random selections.

Why was metacognition less successful in Study 1B? One possibility is that participants essentially selected at random among the estimates throughout Study 1B. Participants might have had to decide randomly if the numerical cues were too difficult to reason about (in comparison to the verbal stimuli in Study 1A) or if the three estimates were similar enough that participants had little basis for determining at the item level which was most accurate.

But another hypothesis is suggested by the fact that participants in Study 1B were actually numerically *worse* than random performance and that they exhibited a numerical preference for the *less* accurate of the initial estimates. The item-based judgments decisions may have been led astray by other, misleading cues. As reviewed previously, item-based judgments can be erroneous when a judge's perception of an item is systematically influenced by variables unrelated to the judgments being made. Indeed, there was evidence for just such a bias: participants relied too much on their more recent estimate. This tendency is erroneous because, as noted above, first estimates were more accurate than second estimates. However, participants in Study 1B showed exactly the opposite pattern in their final responses: they were less apt to choose their first estimate ($M = 23\%$) than their second estimate ($M = 34\%$), $t_{(50)} = -2.54$, $p < .05$, 95% CI: [−19%, −2%], which would systematically increase the error of their reports. One reason for this pattern may be that the second guess was made more recently (indeed, it was made immediately before the final selection phase) and thus the knowledge sampled in that response was closer to what was active at the time that participants made the final selection. Participants may have also been more apt to explicitly remember their experience entering the second estimate than the first and thus favored the estimate that they remembered making.

The hypothesis that participants were misled by their own personal experience when making item-based decisions predicts that individuals with a different subjective experience might be able to more effectively decide among the same set of estimates. We tested this hypothesis in Study 2 by exposing the same options to a new group of decision-makers.

Study 2

In Study 2, we tested whether item-based decisions between three numerical estimates are *always* difficult, or whether the participants in Study 1B were additionally being misled by their subjective experience. We asked a new set of participants to decide between the estimates (and the average of those estimates) made by participants in Study 1B. Each participant in Study 2 completed the same initial estimation phases, but rather than decide among the three numbers represented by their own first, second, and average estimate, they decided among the estimates of a Study 1B participant to whom they were randomly yoked (see Harvey & Harries, 2003, for a similar procedure applied to between-person aggregation).

This study presents participants with the same alternatives to decide among, but with a different prior experience. Participants in Study 2 had made a different set of original estimates, presumably based off an idiosyncratically different base of knowledge than the original participant to whom they were yoked. For these new participants, none of the final options is likely to represent an estimate they just made. Thus, Study 2 can tease apart two accounts of why the original participants' judgments in Study 1B were no better than chance. If the three estimates were inherently difficult to discriminate in item-based judgments or given numeric cues, then the new participants should show similar difficulties. If, however, the participants in Study 1B were additionally hampered by how the response options related to their past experience and knowledge—such as the fact that one of the options represented an estimate that they had just made—then new participants with a different knowledge base might more effectively decide among the same set of estimates.

Method

Participants

Forty-six people participated in Study 2, each of whom was randomly yoked to one of the first 46 participants run in Study 1B.

Procedure

Participants initially made their own first and second estimates following the procedure of the prior studies. In each phase, participants saw the questions in the same order as the Study 1B participant to whom they were yoked. The final decision phase also followed the same procedure as in Study 1B, except that the three response options for each question were no longer the values of the participant's own first, average, and second estimates; rather, they were the three values of the Study 1B participant to whom the current participant was yoked. Participants in Study 2 saw the same instructions as participants in Study 1B, which referred only to *a multiple-choice decision between three possible answers*.

Results

Accuracy of estimates

As in prior studies, the first estimates ($M = 588$, $SD = 371$) made by the Study 2 participants had lower error than their second estimates ($M = 649$, $SD = 428$), although this difference was only marginally significant, $t_{(45)}$=-1.67, $p = .10$, 95% CI: $[-135, 13]$. Again, even the first estimate was numerically outperformed by the average ($M = 560$, $SD = 349$). This effect was not reliable when considering just the Study 2 participants, $t_{(45)} = -1.61$, $p = .11$, 95% CI: $[-62, 7]$; as the initial estimation phases were identical between Study 1 and Study 2, we attribute this lack of significance to the reduced power of the smaller sample in Study 2. (In an analysis presented later in the General Discussion, we pooled the initial estimation phases, which never varied across studies, and found a robust benefit of averaging the two estimates.) Note, however, that these initial estimates were never actually seen in the final decision phase of Study 2. Rather, participants in Study 2 decided among the first, average, and second estimate of a participant from Study 1B to whom they had been yoked.

Importantly, these yoked participants' initial estimates differed from the new participants' initial estimates. On 90% of trials, the second estimate made by the new, Study 2 participant did not match either of the yoked Study 1B participant's estimates; indeed, on 79% trials, *neither* of the new participants' estimates matched either of the original estimates. Thus, when presented with the yoked Study 1B participant's estimates in the final decision phase, the new participants were viewing a novel set of estimates and could not, for instance, adopt a strategy of selecting their second, more recent estimate. Below we describe the consequences of this for participants' strategy selection and for the accuracy of the selected estimates.

Final selections

Although the new Study 2 participants saw the same response options as the Study 1B participants who originally provided the estimates, the Study 2 participants did not share the same erroneous preference for the second estimate over the first estimate. Recall that in Study 1B, participants were reliably more apt to report their second estimate than their first. This same preference did not obtain among the Study 2 participants viewing the same estimates. In fact, the preference for the second estimate was almost completely *reversed*: the new participants were marginally *less* likely to choose the second estimate ($M = 28\%$, $SD = 16\%$) than the first estimate ($M = 36\%$, $SD = 19\%$), $t_{(45)} = -1.78$, $p = .08$, 95% CI: [−15%, −1%].

Performance of strategies

Because the Study 2 participants were less biased towards the typically inaccurate second estimate, it is plausible that they came closer to the true answers than the original Study 1B participants. Fig. 4 displays the squared error of the responses selected by the Study 2 participants in comparison to the error that would be obtained under the alternate strategies described previously and to the error obtained by the Study 1B participants to whom they were yoked.

Unlike the participants who originally made the estimates, the new participants made selections ($MSE = 442$, $SD = 239$) that resulted in a squared error that was lower (i.e., was more accurate) than what would be obtained by responding completely randomly ($MSE = 510$, $SD = 283$), $t_{(45)} = -3.61$, $p < .001$, 95% CI: [−104, −30]. In fact, the new participants even demonstrated that they were effectively selecting strategies on a trial-by-trial basis. Their estimates had less error than the proportional random baseline *(*$MSE = 489$, $SD = 262$), $t_{(45)} = -3.01$, $p < .01$, 95% CI: [−78, −15], which represents the error that would be obtained if participants had selected the same proportion of trials on a random set of trials. As would be expected from the fact that only the new participants exceeded chance performance, the new Study 2 participants' selections had significantly lower error than those made by the original Study 1B participants to whom they were yoked ($MSE = 513$, $SD = 310$), $t_{(45)} = -2.37$, $p < .05$, 95% CI: [−131, −11].

Discussion

New decision-makers were far more accurate at selecting the most accurate of a first, second, and average estimate than were the judges who originally made those estimates. This result rules out several explanations for the ineffective metacognition observed in Study 1B. Participants in Study 2 saw the same numbers as in Study 1B, in the same display, and in the same order, but were quite successful at deciding among them. Therefore, it was not the case that the numerical estimates were simply too similar to discriminate or that participants are inherently challenged when working with numerical stimuli.

Instead, Study 2 supports the hypothesis that participants in Study 1B were misled by their prior experience with the estimates. Although the numbers in the final decision phase were the same across studies, participants' prior experience with those estimates was not the same: the initial estimates provided by participants in Study 2 generally did not match those of the original participant to whom they were yoked. This differential experience could have altered participants' performance in at least two ways. First, the new participants in Study 2 could have combined their original knowledge with the estimates provided by the original participant, producing the typical benefit of averaging multiple sources of information. However, decision-makers typically underuse such strategies (Bonaccio & Dalal, 2006), so it is not clear that such a strategy would account for all of the gains in Study 2. Indeed, making an initial estimate in response to a question impedes one's later ability to effectively aggregate estimates made by multiple other judges (Harvey & Harries, 2003), indicating that retrieving one's own knowledge does not necessarily improve decisions about others' estimates. Moreover, whatever the contribution of the Study 2 participants' own knowledge, it does not explain why the original Study 1B participants exhibited a reliable but erroneous preference for their second, most recent estimate.

A second, likely critical difference is that only the Study 1B participants had their decisions contaminated by a misleading cue. In Study 1B, participants decided between estimates (and the average of those estimates) that they had just made. These participants exhibited a preference for their more recent estimate over their first estimate, which was inappropriate given that these second estimates were the *least* accurate. Such a preference may have been driven by the recency of the second estimate: participants may have been more apt to recollect entering it and favored it for that reason, or it simply may have been more representative of the subset of their knowledge that participants currently had in mind. By contrast, when the Study 2 participants were presented with the original participants' estimates in the final decision phase, none of the options corresponded to an estimate the decision-makers had just themselves made. These participants exhibited no preference for what was originally the most recent estimate. This pattern is consistent with work (e.g., Benjamin, Bjork, & Hirshman, 1998a; Benjamin et al., 1998b; Jacoby & Whitehouse, 1989; Whittlesea et al., 1990) establishing that irrelevant sources of fluency can mislead judgments: the Study 1B participants appear to have been systematically led astray by the recency or fluency of their most recent estimate, even though such estimates were the least accurate.

Misleading influences of subjective fluency in other domains, such as episodic memory, can be reduced or eliminated when participants are able to attribute the fluency to the correct source (e.g., Jacoby & Whitehouse, 1989; Whittlesea et al., 1990). It is possible, then, that such cues could be less damaging, and perhaps even useful, when used in conjunction with participants' general beliefs about how to decide among multiple estimates. We tested this possibility in Study 3.

Study 3

In Study 3, participants saw *both* the labels (*first guess*, *average*, and *second guess*) and numerical values presented together during the final selection phase. As in Study 1, participants selected among their own estimates, not those of a prior participant.

This combination of cues could result in several patterns of behavior. Participants might respond exclusively on one basis or another. If, for instance, participants relied whenever possible on their general theories about averaging versus choosing, they might perform similarly to the

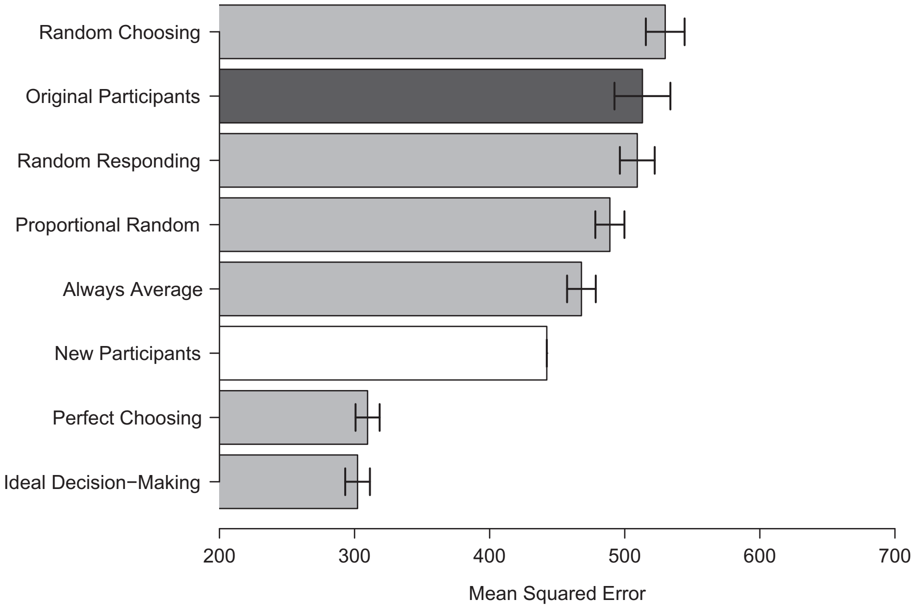

Fig. 4. Mean squared error (MSE) of the final selections made by the new Study 2 participants, in comparison to the MSE obtained by the final selections made by the original Study 1B participants who provided the estimates and to the MSE that would have been obtained under several comparison decision strategies. Error bars indicate the 95% confidence interval of the difference in MSE between the new participants' selections and each comparison mean.

Study 1A participants, who saw only the labels. Conversely, the mere presence of particular estimates that participants had made in the past might be misleading and cause participants to show little evidence for effective metacognition, as in Study 1B. A third possibility is that judges effectively *integrate* theory- and item-level cues. In this case, participants in Study 3 might demonstrate an entirely different—and perhaps better—pattern of performance than participants in either of the prior studies.

Study 3 also included a manipulation of the order of the strategies in the display to assess whether participants' preferences in the prior studies were partially a product of the display.

Method

Participants

Fifty-four people participated in Study 3.

Procedure

The same procedure was followed for the first and second guesses, except that the intervening task was a 15-min language production task. In the third phase, participants were given the same instructions as participants in Study 1A, which explained that they could choose between their first guess, second guess, or average guess and presented examples of each. Participants then viewed the labels from Study 1 presented simultaneously with their actual numerical values (e.g., *Your first guess: 43*).

In Study 3, we also investigated whether the order of the response options in the final decision phase influenced participants' decisions by manipulating this order between participants. Participants were randomly assigned to see the response options either in the order *first guess*, *average*, and *second guess* or the order *first guess, second guess*, *average*; these orders were chosen to vary the order in the display while still retaining the correct temporal order of the first and second estimate. We term the former display the *average-middle display* and the latter the *average-last display*.

Finally, in Study 3, we assessed whether the results of the prior studies were likely to have been influenced by the exclusion of trials in which the two estimates differed by less than 2 percentage points. In Study 3, participants made a final selection for all trials, regardless of the similarity of the estimates. Trials in which the two estimates differ by less than 2 percentage points (19% of trials in Study 3) were still excluded from the primary analysis because they did not include three distinct integer values that participants could choose among. However, because participants actually did make decisions on these trials in Study 3, we also performed a secondary analysis in which all of the trials were included. This analysis revealed that including same-estimate trials only minimally alters the means and does not influence[6] the outcome of any of the critical comparisons; we report the results with the same-estimate trials excluded for consistency with prior experiments.

[6] To understand why these trials had little influence on the results, consider a trial on which the participant's first estimate and second estimate are both 40%. The average of the two estimates is thus 40% as well. Consequently, all three response options in the final decision phase are the same number (40%) and have the same MSE. In such a decision, participants' actual selections necessarily have MSE that is identical to that obtained from selecting randomly, from picking the best of the three estimates, from always averaging, or from any of the other comparison strategies. Thus, these trials do not influence the relative ordering of the participants' decision and comparison strategies.

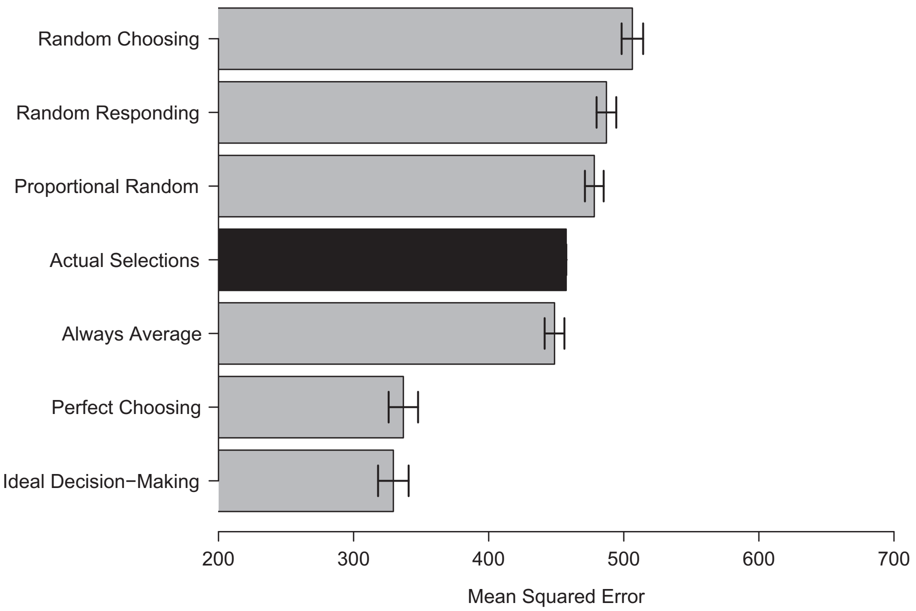

Fig. 5. Mean squared error (MSE) of participants' final selections in Study 3 versus the MSE that would have been obtained under several comparison decision strategies. Error bars indicate the 95% confidence interval of the difference in MSE between participants' actual selections and each alternate strategy.

Results

Accuracy of estimates

As in prior studies, first estimates ($MSE = 504$, $SD = 344$) had somewhat lower squared error than second estimates ($MSE = 543$, $SD = 346$), although this difference was not reliable in Study 3, $t_{(53)} = -1.31$, $p = .19$, 95% CI: [−98,21]. Importantly, however, the average of the estimates ($MSE = 453$, $SD = 303$) had lower error than even the first estimate, $t_{(53)} = -3.09$, $p < .01$, 95% CI: [−84,−18], indicating that an averaging strategy would be effective—if participants applied it.

Final selections

There was no evidence that the rate of averaging differed between the average-middle ($M = 44\%$) and average-last ($M = 47\%$) displays, $t_{(52)} = -0.49$, $p = .63$, 95% CI: [−15%,9%]. Consequently, we collapsed over this variable in the remaining analyses.

Overall, participants reported the average most frequently ($M = 45\%$ of trials, $SD = 22\%$), more than they chose their first guess ($M = 24\%$, $SD = 23\%$) or chose their second guess ($M = 31\%$, $SD = 23\%$). A one-sample *t*-test revealed this rate of averaging was greater than chance, $t_{(53)} = 3.97$, $p < .001$, 95% confidence interval of the mean: [39%,51%].

When participants chose one of the original estimates to report, they chose the more accurate estimate 56% of the time. (Two participants who always averaged were excluded from this analysis.) Recall that, by contrast, the participants in Study 1B were numerically more likely to choose the *less* accurate of the two estimates. Thus, the Study 3 participants, who chose on the basis of both the numerical values and strategy labels, were more accurate in choosing ($M = 56\%$) than the Study 1B participants ($M = 47\%$), who saw the numerical values only. This difference was significant, $t_{(101)} = 2.08$, $p < .05$, 95% CI of the difference: [1%,18%]. Participants' superior choosing accuracy in Study 3 suggests that when the strategy labels were present, participants were less likely to be misled into choosing an inferior estimate.

Performance of strategies

The squared error of participants' actual selections, and the squared error that would have obtained under several alternate strategies, is displayed in Fig. 5.

The combination of labels and numerical values in Study 3 resulted in effective metacognition. The squared error of participants' actual selections ($MSE = 467$, $SD = 305$) was less than what would be obtained by randomly selecting between the three response options ($MSE = 500$, $SD = 318$), $t_{(53)} = -2.90$, $p < .01$, 95% CI: [−57,−10]. In addition, unlike participants in either Study 1A or Study 1B, participants in Study 3 showed evidence for trial-by-trial strategy selection. Actual performance resulted in reliably lower squared error than the proportional random baseline obtained by selecting strategies in the same proportions but on a random set of trials ($MSE = 492$, $SD = 322$), $t_{(53)} = -2.24$, $p < .05$, 95% CI: [−47,−3].

Participants' selections were accurate enough in Study 3 that, unlike in prior studies, their selections did not have reliably greater error than the estimates that would be obtained by simply always selecting the average ($MSE = 453$, $SD = 303$), $t_{(53)} = 1.15$, $p = .26$, 95% CI: [−10,37], although the always-average strategy did still yield numerically better performance. However, participants' selections still resulted in reliably greater squared error than would have been obtained just from choosing with perfect accuracy between the two original estimates (*MSE* = 317, *SD* = 238) and never averaging, $t_{(53)} = 8.75$, $p < .001$, 95% CI: [116,185].

Choosing versus averaging

The above comparison illustrates an important caveat of combining multiple estimates. Averaging the estimates yielded lower squared error than consistently choosing the first estimate or consistently choosing the second estimate, as reviewed above. But participants in all three studies could have made their reporting even more accurate by choosing whichever of the two original estimates was better *on a particular trial*. For example, in Study 3, choosing the better of the two estimates would result in lower squared error than always averaging the estimates, $t_{(53)} = -10.33$, $p < .001$, 95% CI: [−163, −110]. Two characteristics of a decision environment define when choosing can outperform averaging (Soll & Larrick, 2009): (a) the better estimate is substantially more accurate than the worse estimate, and (b) more importantly, the estimates are highly correlated with each other, so that each does not contribute much independent information that could improve the accuracy of the average. The latter is certainly the case for multiple estimates made by the same individual, which are strongly correlated (Herzog & Hertwig, 2009; Vul & Pashler, 2008).

This might suggest that participants would be better served by choosing one estimate rather than averaging them. However, the practical effectiveness of a choosing strategy depends not only on the characteristics of the decision environment, which define the upper bounds of the success of a choosing strategy, but also on the decision-maker's ability to actually identify the better of the two estimates (Soll & Larrick, 2009). This relation is depicted in Fig. 6, which depicts, across all trials, the expected value of a choosing strategy given different probabilities ϕ of identifying the better estimate, as well as the constant squared error resulting from averaging. As described above, in the decision environment of Study 3 (as well as in those of prior studies), always choosing the better estimate ($\phi = 1.0$, *MSE* = 318) yields lower squared error than averaging. However, chance choosing ($\phi = 0.5$, *MSE* = 527) yields greater error than averaging (*MSE* = 456), $t_{(53)} = 7.91$, $p < .001$, 95% CI: [53,88]. The two strategies yield equivalent performance when $\phi = .67$. Thus, participants in the task should have adopted a choosing strategy if they could choose the better estimate two-thirds of the time, but should have otherwise averaged their estimates.

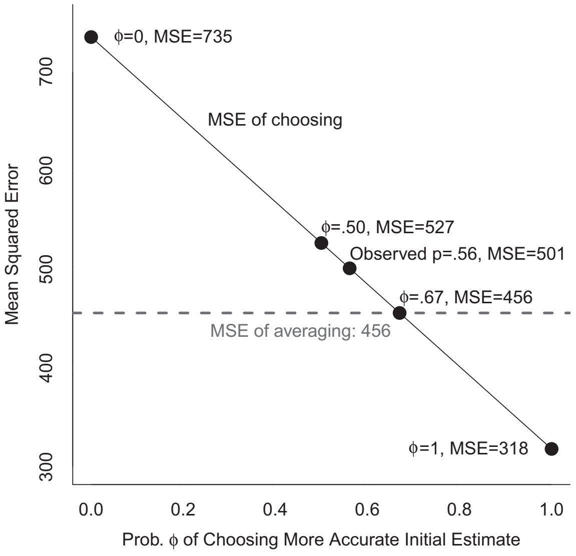

Fig. 6. Mean squared error (MSE) that would be obtained by a choosing strategy, as a function of probability ϕ of choosing the better of the two initial estimates (solid line) and in comparison to the constant MSE obtained by always averaging the estimates (dashed line). Averaging results in lower MSE than choosing unless participants choose the better of the two estimates with probability .67 or greater, which they did not attain.

Can participants realistically obtain this level of choosing accuracy? We again examined the trials on which participants chose one of the original estimates[7] and calculated the proportion *p* of these trials on which participants chose the better of the two original estimates. (Two participants who always averaged were excluded from this analysis.) We compared this *p* to the ϕ that each participant would need, given the particular decision environments they were presented with, to achieve squared error lower than that of a pure averaging strategy. Only 17 of the 52 subjects chose the better original estimate at the rate required for them to outperform a pure averaging strategy. Overall, participants chose the better estimate only 56% of the time, which was well below the rate needed to beat averaging, $t_{(51)} = -2.79$, $p < .01$, 95% CI of the difference: [−17%, −3%]. Given these limits in choosing the better estimate, participants would have been best served by averaging the estimates.

Discussion

The combination of both a cue to a general naïve theory (a strategy label) and item-specific information (the particular numerical estimate yielded by that strategy) resulted in superior metacognitive performance than either basis alone. Compared to participants given only the numerical estimates (Study 1B), participants given both cues were more accurate at identifying the better of their original estimates, and their decisions to report their first, second, or average estimate resulted in significantly lower error than would be expected by chance. Although participants given only the theory-based cues in Study 1A also attained that level of performance, participants in Study 3 addition-

[7] On trials where participants reported the average, it is unknown which of the two original estimates they would have chosen as the better estimate. However, to obtain a *p* any higher than what was estimated from the available data, participants would have to be substantially *better* at choosing on those trials for which they elected *not* to employ a choosing strategy, which seems implausible. Focusing only on trials on which participants actually decided to employ a choosing strategy likely provides an overestimate, if anything, of participants' accuracy in choosing the better original estimate.

ally selected effective strategies on a trial-by-trial basis. Evidence for this comes from the fact that assigning their strategy selections to a random set of trials would have resulted in substantially higher error than was actually observed, indicating that participants had tailored those strategies to the particular trials on which they used them.

Study 3 also provides evidence against two alternate explanations of participants' preferences in the prior studies. First, participants' strategy choices were unlikely to be driven by the location of those strategies in the display, as experimentally manipulating the locations had no effect. Thus, for instance, participants' preference in Study 1B for their second guess cannot be attributed simply to a preference for the last option in the screen because placing the average in that location did not increase the rate at which the average was selected. Second, providing both the theory-level strategy labels and item-level numerical estimates in Study 3 resulted in a pattern of metacognitive performance that was qualitatively different from that observed in our prior studies. This difference suggests that participants given only one of the cues in prior studies were not using it to retrieve the other (e.g., retrieving the numerical estimates associated with the labels *first guess* and *second guess*) and responding on the basis of both, which should have emulated the performance observed in Study 3.

Thus, Study 3 demonstrates that metacognitive decisions about how to combine multiple estimates can be made most effectively when both theory-level and item-level bases for those decisions are available. Nevertheless, although Study 3 yielded more successful metacognition than the prior studies, it also revealed considerable limitations. Participants could have reported more accurate answers had they been able to choose the better of the two original estimates with a high level of success. However, an examination of trials on which participants chose one of the original estimates indicated that participants were not successful enough at identifying the better estimate to make a choosing strategy effective. From this perspective, participants' preference for aggregating estimates was an appropriate hedge against the inability to choose the better estimate.

General discussion

Four studies investigated how individuals made metacognitive decisions related to multiple estimates. Participants made two initial estimates, on different occasions, of the answers to world knowledge questions. In a final phase, they decided whether to report the average of their estimates or one of the original estimates as their final, most accurate answer.

Replicating past results, the average of two estimates made on different occasions was more accurate than either of the individual estimates. Because the initial estimation phases were identical across studies, we pooled participants from all four studies ($N = 213$) to assess the comparative accuracy of the initial estimates. First estimates ($MSE = 537$, $SD = 338$) had lower squared error than second estimates ($MSE = 606$, $SE = 376$), $t_{(212)} = -3.82$, $p < .001$, 95% CI: [−105, −34], but the average of the two estimates ($MSE = = 500$, $SE = 312$) had even lower error than the first, $t_{(212)} = -4.27$, $p < .001$, 95% CI: [−55, −18]. This replicates the benefit of averaging multiple estimates from the same individual (Herzog & Hertwig, 2009; Rauhut & Lorenz, 2010; Vul & Pashler, 2008) and demonstrates that the second estimates contributed new information not incorporated into the first estimate. The focus of our study, however, was whether participants would report the combined estimate or one of the original estimates as their actual final answer.

Across the four studies, the cues available in the final decision phase were manipulated to emphasize theory-based decisions, item-based decisions, or both. In Study 1A, participants were provided descriptions of the sources of the estimates (i.e., *first guess*, *second guess*, *average*) but no information about the specific numerical estimates those sources yielded on a particular trial. These participants exhibited an overall preference for the strategy that minimized error—averaging—but showed no evidence of being able to choose which option would be most effective for a particular trial. In Study 1B, participants were given only item-level cues—numerical values—and no information about what yielded the numbers. These participants performed no better than randomly selecting which value to report. This lack of metacognitive effectiveness in item-level judgments was unlikely to be due just to the difficulty of discriminating between similar numerical estimates. Rather, participants appear to have been systematically misled by their preference for their most recent estimate, which was actually the least accurate estimate. This interpretation was supported by Study 2, in which new participants were given the same values, but without the experience of having made one of those estimates more recently than the other; these participants were far more successful at reporting accurate estimates. Finally, in Study 3, combining the labels from Study 1A with the numerical values from Study 1B yielded the best metacognitive performance. Not only did participants generally prefer the best overall strategy (averaging), they also showed evidence of selecting the most effective strategy on a trial-by-trial basis. Below, we discuss the implications of these results for theories of how decision-makers make use of multiple estimates or cues, particularly those stemming from multiple judges.

To combine or to choose?

When faced with multiple cues to a decision, such as several different estimates, decision-makers can either choose a single cue (Gigerenzer & Goldstein, 1996) or attempt to combine cues. Combining estimates, either from the same individual or different individuals, can improve judgment accuracy by reducing the influence of random error and of bias (Yaniv, 2004). When the estimates are sufficiently independent (i.e., the errors are not correlated), and one judge is not substantially more accurate than another, the average can outperform even choosing the *best* judge or cue (Soll & Larrick, 2009). When the estimates are less independent, such as when they come from the same judge, averaging produces smaller benefits (Herzog

& Hertwig, 2009; Rauhut & Lorenz, 2010; Vul & Pashler, 2008) and can be outperformed by choosing the most accurate judge.

The present study represented the latter type of environment. In most cases, the better of participants' original estimates was closer to the true answer of the question than was the average of those estimates. This is to be expected. The average only outperforms both original estimates on trials in which the two estimates bracket the true answer, and bracketing is relatively rare when estimates are as strongly correlated as are two estimates made by the same individual with only a short delay in between. In principle, then, choosing the better original estimate should outperform averaging. However, an examination of participants' attempts to implement a choosing strategy indicated that they were scarcely better than chance at identifying the better of the two estimates. Given these limits, it is actually averaging that would have resulted in lower error. This analysis reveals the important constraints provided by the abilities of the decision maker: even in decision environments in which a choosing strategy hypothetically *could* outperform averaging, averaging may be more effective if participants cannot choose the appropriate cue. (Note, however, that combining multiple cues may have other disadvantages, such as the need to retrieve multiple cues from memory; Gigerenzer & Goldstein, 1996.) In light of these constraints, participants' preference for the average appears appropriate.

The use of an apparently suboptimal strategy as a hedge against the inability to execute a hypothetically superior strategy can also be seen in other cognitive domains. For example, episodic memories can be more easily retrieved in contexts similar to the ones present at learning (Tulving & Thomson, 1973). However, learners rarely know the exact circumstances under which they will later need to use information, so studying information with a variety of contexts or cues can be a beneficial hedge (Finley & Benjamin, 2012).

Analytic and nonanalytic bases for judgment

How did participants decide whether or not to average their estimates? It has frequently been suggested (e.g., Kelley & Jacoby, 1996; Koriat, 1997; Kornell & Bjork, 2009) that metacognitive decisions may be made on multiple bases. The present work supported this hypothesis and extended it to the domain of combining multiple estimates. As described above, participants' success at identifying the most accurate estimate varied depending on whether the cues in the environment were likely to support a judgment based on a naïve theory or based on item-level characteristics. In Study 1A, participants saw only descriptions of how particular estimates were generated (e.g., the participant's first estimate, or the average of the two estimates), which were likely to support decisions based on participants' general beliefs about the effectiveness of the labeled strategies. In this case, participants displayed some evidence for successful metacognition; the estimates they selected as their final reports exhibited lower error than what would be obtained under chance selection. By contrast, in Study 1B and in Study 2, participants saw no overt cue to naïve theories about the value of averaging versus choosing. Rather, they received only the numeric estimates produced by each strategy. In this case, we expected participants' judgments were more likely to be based on an item-specific judgment of how plausible each of those estimates was as an answer to the question. Differences in such plausibility may stem from differences in what subset of knowledge is currently active or sampled by participants or from participants' ability to remember making some estimates but not others. Given only these item-level cues, participants exhibited no reliable evidence for effective metacognition; their final reports were no better than what would be obtained by selecting randomly between the estimates. This discrepancy reveals how the quality of decision-making can vary depending on what basis for judgment is supported by the environment.

One important distinction is that, although past work has often contrasted theory level versus item-level judgments (e.g., Kornell et al., 2011) or analytic versus nonanalytic reasoning (Jacoby & Kelley, 1987; Kelley & Jacoby, 1996) as competing influences, the present work suggests that individuals can productively integrate multiple kinds of cues. Participants' metacognition was most effective in Study 3, in which they were presented with both the strategy labels and the numerical values. Only in that study did participants show reliable evidence of selecting the appropriate strategy on a trial-by-trial basis. This result indicates that different cues to judgment are not mutually exclusive; rather, decision-makers are able to combine them to achieve qualitatively superior metacognition.

Misleading effects of subjective fluency

Although the item-level numerical estimates were beneficial when combined with the strategy labels in Study 3, they did not support successful metacognition when presented alone in Study 1B. Participants in that study frequently chose their more recent estimate as their final report, even though it was on average the *least* accurate estimate. This systematically erroneous preference indicates that the challenge of Study 1B was not just due to the difficulty of selecting among three similar numerical values. Rather, it suggests decisions were misled by participants' recent experience making the estimates. The second estimate was the one that participants made most recently, and it may have seemed especially truthful or plausible because it was more consistent with participants' present state of mind or because they could remember making the judgment in the earlier phase. This recency hypothesis receives further support from Study 2, in which new participants, for whom none of the values represented one of their recent estimates, were far more effective at selecting among the same values.

This result is consistent with a large literature indicating that judgments and decision-making are influenced by processing fluency or effortfulness, as determined by factors such as recency and ease of perception. (For review, see Alter & Oppenheimer, 2009; Benjamin & Bjork, 1996; Oppenheimer, 2008.) In short, stimuli that can be easily or quickly processed are better liked and are believed to be previously encountered and more accurate (Alter &

Oppenheimer, 2009; Benjamin & Bjork, 1996). Making decisions on the basis of fluency is likely to be an effective heuristic overall because the factors that determine subjective fluency typically do relate to the objective properties being judged; for example, it is rational to judge an easy-to-process item as previously encountered because previously viewed items typically *are* easier to process than new items (Benjamin et al., 1998a). However, the effectiveness of the heuristic breaks down in situations in which subjective fluency is systematically influenced by factors unrelated to, or inversely related to, the property being judged, as in the present case.

In particular, fluency leads to overreliance on one's present state of knowledge when attempting to infer others' knowledge or one's own future or past knowledge. For example, in episodic memory, learners underestimate both how much they can learn and how much they will forget in the future (the *stability bias*; Kornell & Bjork, 2009, see also Koriat et al., 2004); it has been argued (Kornell et al., 2011) that this bias reflect an overreliance on the present ease or difficulty of processing an item and an underuse of naïve theories about learning and forgetting. Similarly, in conversation, speakers and hearers appear to overrely on their own knowledge, rather than that of their interlocutor, in producing and understanding language (Brown-Schmidt & Hanna, 2011; Keysar, Barr, Balin, & Brauner, 2000). This overreliance on one's current knowledge is consistent with the current participants' frequent use of their most recent estimate, even though that estimate was generally the least accurate. A similar difficulty in overcoming the influence of one's own perspective may also explain why participants are reluctant to adopt the judgments of others or aggregate them with their own (Bonaccio & Dalal, 2006), as we discuss below.

Although subjective fluency stemming from irrelevant sources can mislead judgments, these errors can often be reduced or eliminated when fluency can be attributed to its proper source (for review, see Alter & Oppenheimer, 2009). For instance, although visual clarity influences whether participants judge a word as previously encountered, these effects vanish if participants are told in advance about the manipulation and can attribute the variance in fluency to its proper source (Whittlesea et al., 1990; see also Jacoby & Whitehouse, 1989). In Study 3, pairing the numerical estimates with labels describing their sources could have helped participants correctly attribute the fluency of the second estimate to its recency rather than to its accuracy as an answer to the world knowledge question, thus reducing the misleading bias to report the second estimate. Indeed, the combination of cues allowed participants not only to improve their metacognition, but to achieve performance superior to that of either cue presented alone. This result speaks to the ability of proper attributions to override, and even reverse, misleading effects of fluency.

Sources of recency effects

The evidence reviewed above suggests that, given only the numerical values of judgments they had made previously, decision-makers are inaccurately biased to report judgments they made more recently. Several variables may contribute to this bias. If contextual factors lead participants to randomly sample a proper subset of their knowledge at any given time, recently made estimates are likely to draw on a subset of knowledge more similar to the decision-maker's present state of mind than an estimates made in a more distant context. In addition, participants are likely to be more apt to consciously remember making a more recent estimate than an earlier one, and recollecting this experience may also contribute to the feeling that the second estimate is more plausible. Additionally, participants may prefer to report an estimate they can remember making previously so that their judgments appear consistent. Any or all of these factors may have contributed to the preference for the recent estimate in the numbers-only condition, and an interesting avenue for future work would be to examine which sources of evidence underlie these item-based decisions and to what degree.

Metacognition about multiple estimates

When faced with multiple possible answers to a question, how should they be chosen among or combined? In the present study, as well in the work of Müller-Trede (2011), participants were faced with such a decision because they had provided multiple answers to each question. But similar decisions also arise when decision-makers are given estimates from multiple judges or when an advisor provides advice that differs from one's own perspective. The strategies and success of participants deciding among several of their own estimates, then, can also inform broader accounts of how decision-makers use multiple, conflicting judgments.

In particular, participants' decisions about how to combine several self-generated estimates appear strikingly similar to what prior studies have observed about their decisions about how to combine estimates from several different people. There are at least two parallels. First, decision-makers sometimes combine estimates but do so with suboptimal frequency. Although participants presented with the opportunity to use several judges' estimates sometimes average them, they often choose one judge's estimate even where averaging would be beneficial (Soll & Larrick, 2009), and they rely too heavily on their own estimate (Bonaccio & Dalal, 2006). Similarly, in the present studies, participants presented with multiple self-generated estimates underused averaging and instead relied too heavily on choosing their second estimate. The second parallel is that assessments of decision-makers' naïve theories about averaging reveal only a weak appreciation for averaging. When asked to explicitly reason about combining the estimates of multiple judges, only a bare majority of participants, or even slightly fewer, correctly appreciate that averaging several judges can outperform the average judge (Larrick & Soll, 2006; Soll, 1999). Analogously, in the present study, participants given just descriptions of the strategies only slightly preferred the average over their first estimate or their second estimate.

The similarity of participants' behavior in combining their own estimates over time and in combining the estimates of multiple judges suggest a common basis to both judgments—and places important constraints on what that basis might be. Some past theories have attributed underuse of others' judgments to social factors, such as a belief that one is a more skilled judge than others (Harvey & Fischer, 1997). (For further discussion of such accounts, see Bonaccio & Dalal, 2006; Krueger, 2003.) The present studies suggest that such factors cannot be the only reason decision-makers do not aggregate estimates: even when all the estimates were self-generated, participants still underused a strategy of combining estimates.

Other theories (e.g., Harvey & Fischer, 1997; Harvey & Harries, 2003; Lim & O'Connor, 1995) have attributed participants' decisions about using multiple estimates, and in particular their underuse of others' advice, to a primacy preference. Judges have already formed their own opinions, so when they receive another estimate from an advisor, they are reluctant to alter their original preference. Thus, it is the fact that one's opinion comes first, rather than the fact that it is self-generated, that causes it to be overweighted. This theory effectively accounts for the typical judge-advisor experiment, in which judges make their own initial estimate before receiving the estimate from the advisor (Bonaccio & Dalal, 2006). However, in the present studies, both the initial and later estimate were self-generated, deconfounding primacy from one's own viewpoint. In these cases, participants chose their *recent* guess more than their initial one, and their decision accuracy suffered as a result. This result indicates that ineffective use of multiple estimates is not always driven by a primacy preference; indeed, sometimes the exact reverse preference obtains. In addition, decision-makers overweight their own opinion even when it is formed after advice is given (Yaniv & Choshen-Hillel, 2012).

Why, then, are decisions about using multiple estimates often made suboptimally? The present study suggests two factors that influenced decision-makers' behavior both in the present and prior studies. First, decision-makers often hold incorrect beliefs about the most effective strategy. Participants appear to have incorrect naïve theories about the mathematical benefits of averaging (Soll, 1999), and when asked in the present experiments to decide on the basis of strategy descriptions, only weakly preferred the most effective strategy. Second, a cognitive constraint common to both between-person and within-person aggregation—and, indeed, many other tasks reviewed above—is the difficulty of overcoming one's present perspective. Both in deciding between one's current estimate versus a prior estimate and in deciding between one's own estimate versus another individual's, decision-makers appear to rely too heavily on their present state of mind. They choose their current estimate over a past one, and their own estimate over another person's. The fact that participants given no cues to a general naïve theory, who likely had to respond based only on item-level fluency or plausibility, fared no better than chance performance suggests that this latter constraint on decision-making may be a particularly pernicious one. This account is similar to the hypothesis (Yaniv, 2004) that decision-makers overweight their own opinion because they have internal access to the evidence supporting their own judgments but not others'. However, our account emphasizes that differential feelings of fluency or accessibility need not arise only from a self-versus-other distinction. Judgments, including multiple self-generated estimates, may be closer to or further from one's present state of mind for multiple reasons. This broader proposal can account for how—and how effectively—decision-makers use multiple estimates both in the current and past studies.

One caveat in concluding that decision-makers insufficiently value combining multiple estimates is that the present participants were presented with a task in which the benefits of doing so were relatively modest. As noted above, averaging multiple estimates produces larger gains in accuracy when estimates are more independent (less correlated) than are estimates made from the same individual. Participants may have been more apt to recognize the value of averaging had it yielded larger gains in accuracy (Larrick & Soll, 2006). Nevertheless, even in the present task, averaging still conferred a benefit over using the first estimate alone or second estimate alone, and participants could have taken advantage of this benefit more than they actually did. In fact, participants were presented with a decision environment that likely *encouraged* averaging by eliminating many of the typical barriers to implementing such a strategy. Metacognitive strategies may be less commonly implemented when they must be self-initiated than when they are supported by external cues (e.g., Craik, 1983; Tullis & Benjamin, 2012). Moreover, a disadvantage of integrating multiple cues is that it might be time-consuming to retrieve and integrate all of the needed information (Gigerenzer & Goldstein, 1996; Harvey & Fischer, 1997; but see Lim & O'Connor, 1995, for evidence that this is not the primary reason judges underuse averaging). However, in the present study, the average was present in the environment, eliminating the need for participants to perform any time-consuming operations or initiate the strategy on their own. Thus, the underuse of averaging despite such aids likely reveals a genuine underappreciation of its value.

Conclusion

Judgments can be improved by considering multiple estimates. Even when the estimates under consideration are all self-generated, averaging them would allow decision-makers to harness the crowd within and improve judgment accuracy by reducing the random error of their estimates. Although in principle averaging produces greater error than always identifying the better estimate, participants are often not particularly skilled at identifying the better estimate, making averaging the more advisable strategy.

However, being faced with multiple estimates also requires a decision about how to use those estimates. Although decision-makers make some attempt to combine estimates generated by different individuals, they often do so suboptimally (Bonaccio & Dalal, 2006). Similarly, participants in the present study displayed some preference for the normatively most effective strategy—averaging—but generally underused it. In particular, the efficacy of participants' judgments depended on whether the cues at the time of the decision favored a decision based on participants' general naïve theory or on item-level judgments.

Participants preferred the average when given explicit descriptions of the strategies, but appeared misled by the recency of their second estimate when the task favored item-based decisions. Metacognition was at its most effective when both cues were present; only with both cues did participants show evidence of adopting the most effective decision strategy on a trial-by-trial basis.

These results highlight the ability of decision-makers to select decision strategies on a per-decision basis, and they demonstrate that theory-level and item-level bases for judgment can be productively combined to qualitatively enhance metacognitive decision-making. Further, they suggest that the difficulty of making effective use of multiple estimates is not driven purely by social differences between one's self and one's advisors. Rather, more general difficulties in forming effective naïve theories and in overcoming the misleading influence of one's current perspective can keep decision-makers from fully harnessing the power of multiple judgments.

Acknowledgments

This work was supported by National Science Foundation Graduate Research Fellowship 2007053221 and a traineeship from National Institutes of Health Language Processing Training Program 5T32HD055272-13 to Scott H. Fraundorf. We thank the members of the Human Memory and Cognition Lab for their comments and suggestions, and Ran He, Tuan Lam, Euil Lim, Molly Lewis, Sujin Park, Samantha Schaller, Kristen Tooley, and Alison Trembacki for data collection.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*, 219–235. http://dx.doi.org/10.1177/108868309341564.

Benjamin, A. S., Bjork, R. A., & Hirshman, E. (1998a). Predicting the future and reconstructing the past: A Bayesian characterization of the utility of subjective fluency. *Acta Psychologica, 98*, 267–290. http://dx.doi.org/10.1016/S0001-6918(97)00046-2.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Mahwah, NJ: Erlbaum.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998b). The mismeasure of memory: When retrieval fluency is misleading as a metacognitive index. *Journal of Experimental Psychology; General, 127*, 55–68. http://dx.doi.org/10.1037/0096-3445.127.1.55.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Performance, 101*, 127–151. http://dx.doi.org/10.1016/j.obhdp.2006.07.001.

Brown-Schmidt, S., & Hanna, J. E. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialogue & Discourse, 2*, 11–33. http://dx.doi.org/10.5087/dad.2011.102.

Central Intelligence Agency (2011). *The world factbook*. <https://www.cia.gov/library/publications/the-world-factbook/>.

Craik, F. I. M. (1983). On the transfer of information from temporary to permanent memory. *Philosophical Transactions of the Royal Society of London, Series B, 302*, 341–359. http://dx.doi.org/10.1098/rstb.1983.0059.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13*, 171–192. http://dx.doi.org/10.1016/0030-5073(75)90044-6.

Finley, J. R., & Benjamin, A. S. (2012). Retrieval cue variability: When and why are two meanings better than one? (submitted for publication).

Friedman, M. P., Burke, C. J., Cole, M., Keller, L., Millward, R. B., & Estes, W. K. (1964). Two-choice behavior under extended training with shifting probabilities of reinforcement. In R. C. Atkinson (Ed.). *Studies in mathematical psychology* (Vol. 9, pp. 250–316). Stanford, CA: Stanford University Press.

Galton, F. (1907). Vox populi. *Nature, 75*, 450–451. http://dx.doi.org/10.1038/075450a0.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669. http://dx.doi.org/10.1037/0033-295X.103.4.650.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes, 70*, 117–133. http://dx.doi.org/10.1006/obhd.1997.2697.

Harvey, N., & Harries, C. (2003). Effects of judges' forecasting on their later combination of forecasts for the same outcomes. *International Journal of Forecasting, 20*, 391–409. http://dx.doi.org/10.1016/j.ijforecast.2003.09.012.

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

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

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1–55. http://dx.doi.org/10.1016/0010-0285(92)90002-J.

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 from estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1068–1074. http://dx.doi.org/10.1037/a0019694.

Jacoby, L. L., & Kelley, C. M. (1987). Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin, 13*, 314–336. http://dx.doi.org/10.1177/0146167287133003.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General, 118*, 126–135. http://dx.doi.org/10.1037/0096-3445.118.2.126.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35*, 157–175. http://dx.doi.org/10.1006/jmla.1996.0009.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*, 32–38. http://dx.doi.org/10.1111/1467-9280.00211.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609–639. http://dx.doi.org/10.1037/0033-295X.100.4.609.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370. http://dx.doi.org/10.1037/0096-3445.126.4.349.

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

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processing. *Journal of Experimental Psychology; General, 153*, 643–656. http://dx.doi.org/10.1037/0096-3445.133.4.643.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*, 449–468. http://dx.doi.org/10.1037/a0017350.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*, 787–794. http://dx.doi.org/10.1177/0956797611407929.

Krueger, J. I. (2003). Return of the ego—Self-referent information as a filter for social prediction: Comment on Karniol (2003). *Psychological Review, 110*, 585–590. http://dx.doi.org/10.1037/0033-295X.110.3.585.

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

Lim, J. S., & O'Connor, M. (1995). Judgemental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making, 8*, 149–168. http://dx.doi.org/10.1002/bdm.3960080302.

Lovie, A. D., & Lovie, P. (1986). The flat maximum effect and linear scoring models for prediction. *Journal of Forecasting, 5*, 159–168. http://dx.doi.org/10.1002/for.39850050303.

Mannes, A. E. (2009). Are we wise about the wisdom of crowds? The use of group judgments in belief revision. *Management Science, 55*, 1267–1279. http://dx.doi.org/10.1287/mnsc.1090.1031.

Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of indivdiuals or crowds? *Cognitive Science, 32*, 1133–1147. http://dx.doi.org/10.1080/03640210802353016.

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

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*, 237–241. http://dx.doi.org/10.1016/j.tics.2008.02.014.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 534–552.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *IRE Professional Group on Information Theory PGIT-4*, 171–212.

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

Rauhut, H., & Lorenz, J. (2010). 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*, 191–197. http://dx.doi.org/10.1016/j.jmp.2010.10.002.

Rawson, K. A., Dunlosky, J., & McDonald, S. L. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology, 55A*, 505–524. http://dx.doi.org/10.1080/02724980143000352.

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology, 38*, 317–346. http://dx.doi.org/cogp.1998.0699.

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

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*. New York: Random House.

Tullis, J. G., & Benjamin, A. S. (2012). The effectiveness of updating metacognitive knowledge in the elderly: Evidence from metamnemonic judgments of word frequency. *Psychology and Aging, 27*, 683–690. http://dx.doi.org/10.1037/a0025838.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 353–373. http://dx.doi.org/10.1037/h0020071.

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

Welsh, M. B., Lee, M. D., & Begg, S. H. (2009). Repeated judgments in elicitation tasks: Efficacy of the MOLE method. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 64–70). Austin, TX: Cognitive Science Society.

White, C. M., & Antonakis, J. (2013). Quantifying accuracy improvements in sets of pooled judgments: Does dialectical bootstrapping work? *Psychological Science, 24*, 115–116. http://dx.doi.org/10.1177/0956797612449174.

Whittlesea, B. W. A., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language, 29*, 716–732. http://dx.doi.org/10.1016/0749-596X(90)90045-2.

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

Yaniv, I., & Choshen-Hillel, S. (2012). Exploiting the wisdom of others to make better decisions: Suspending judgment reduces egocentrism and increases accuracy. *Journal of Behavioral Decision Making, 25*, 427–434. http://dx.doi.org/10.1002/bdm.740.