Trying again is overrated:

Reconsidering quantitative estimates does not improve accuracy, but people think it does

Julia A. Minson

Harvard Kennedy School of Government

Daniella M. Kupor

Boston University

Bradley R. DeWees

United States Air Force

ABSTRACT

People commonly advise others to reconsider their initial judgments before finalizing them in order to achieve greater accuracy. The current research investigates the wisdom of this advice. In fourteen experiments conducted across a diversity of estimation contexts, incentive structures, elicitation instructions, and estimator populations, we find that reconsidering a quantitative judgment does not seem to increase its accuracy. This failure to improve persists even when people are highly incentivized to do so, a phenomenon which defies the expectations of both lay people and experts. We uncover observers' rationale for their misguided expectations and find that manipulating this rationale attenuates the misprediction. In so doing, we shed new light on the psychology of quantitative judgment and provide novel insight into the managerial consequences of believing that trying again improves judgment accuracy.

Keywords: estimation accuracy, judgment under uncertainty, lay beliefs

'Tis a lesson you should heed:
Try, try, try again.
If at first you don't succeed,
Try, try, try again.
~~ British Proverb

Quantitative judgments about market conditions, product performance, and human capital often guide firms' decisions. As a result, managers and their employees expend substantial resources in their attempts to formulate accurate judgments and forecasts (Hamlett 2020; Ladesma 2004; Tetlock, Mellers, Rohrbaugh, and Chen 2014; Zak and Waddell 2010). In these endeavors, as in many others, people often adhere to the folk wisdom in the proverb above by trying again and reconsidering their initial judgments, hoping that judgment accuracy will improve. Thus, financial analysts are encouraged to revisit analyses they have previously completed before finalizing them (Spiech 2005), managers are encouraged to revisit their strategic decisions before implementing them (Shaw 2019; Slater 2010; Stevenson 2018; Villamor 2020), and consumers are advised to improve their decision-making by reconsidering the decisions they have made before setting them in stone (Horner 2020; Singh 2019).

In the present manuscript, we explore the accuracy of this widely held assumption and its implications. We do this by examining the actual accuracy improvement that results from reconsidering a judgment under a wide variety of conditions and by contrasting the results of reconsideration with the expectations of observers. Although reconsidering quantitative judgments does not appear to lead to any consistent improvement, observers expect it to. We then document the implications of the disconnect between judgment performance and observer expectations for common managerial decision-making problems.

Related Prior Research

A growing literature has examined the accuracy benefits of averaging a person's multiple estimates (Ariely et al. 2000; Gaertig and Simmons, 2020; Herzog and Hertwig 2009, 2014a, 2014b; van Dolder and van den Assem 2018; Vul and Pashler 2008). This research on the accuracy of the "crowd within" has examined the conditions under which the average of two estimates from the same judge outperforms the accuracy of the initial estimate. However, this work has not focused on the *relative* accuracy of first versus second estimates. This is an important gap because making a second estimate is a strategy that decision-makers often endorse.

Although the papers in the "crowd within" tradition instructed participants to make two estimates of the same unknown quantity in order to calculate the benefits of averaging, most do not report inferential statistics regarding the relative accuracy of the first versus the second estimates (e.g. Ariely et al. 2000; Barneron, Allalouf and Yaniv, 2019; Fujisaki, Honda and Ueda, 2018; Hourihan and Benjamin, 2010; Kim, 2016; Muller-Trede, 2011; Rauhut and Lorenz, 2010; White and Antonakis, 2013). The papers that do report these statistics find inconsistent results. For example, Herzog and Hertwig (2009, 2014) report that second estimates in their data were of approximately equal accuracy to first estimates in some conditions, but slightly more accurate in others. By contrast, Fraundorf and Benjamin (2014), Krueger and Chen (2014), Steegen, Dewitte, Tuerlinckx and Vanpanemel (2014), and Vul and Pashler (2008) found second estimates to be less accurate, although these results were not uniformly reliable.

Examining the broader judgment and decision-making literature at first blush suggests that second estimates may on average be more accurate than first estimates. First, to the extent that revisiting a judgment results in additional cognitive processing, it should improve accuracy. Indeed, substantial literature reveals that devoting additional cognitive effort and deliberation to judgments increases their accuracy across a wide range of estimation domains (e.g., Maki et al. 1990; Mata, Ferreira, and Sherman 2013; Moxley, Ericsson, Charness, and Krampe 2012; Obrecht and Chesney 2018; Pennycook and Rand 2019). This is in part because greater deliberation can correct initial judgment errors (Sah, Malaviya, and Thompson 2018), lead people to incorporate a wider expanse of available information (and thus reduce reliance on heuristics; Obrecht and Chesney 2018), and provide the cognitive resources necessary to correctly carry out any necessary calculations (Logie, Gilhooly, and Wynn 1994). As a result, people's common failure to engage in sufficient deliberation is theorized to underlie many non-normative judgments and decisions (Kahneman and Tversky 1972; Obrecht and Chesney 2018). To the extent that reconsidering an initial estimate prompts greater deliberation, this literature suggests that reconsideration should improve accuracy.

Related research suggests that greater deliberation can also increase people's likelihood of noticing task-relevant cues that are necessary for accurate judgment formulation. For example, greater deliberative processing prompts more thorough visual search of immediately available task-relevant information and of external (non-immediately visible) task-relevant information, both of which can improve judgment quality (Horstmann, Ahlgrimm, and Glöckner 2009; Levin, Huneke, and Jasper 2000; Taber 2011). Potentially as a result, interventions that limit people's ability to engage in additional processing (e.g., time pressure and cognitive load) often reduce judgment accuracy (ALQahtani et al. 2016; Ziv and Leiser 2013).[1]

Lay intuition seems to wholeheartedly endorse the benefits of the "try, try again" approach when it comes to judgment accuracy. Indeed, people employ an accuracy-effort framework, in which they decide whether to tradeoff greater cognitive effort in order to achieve greater accuracy (Payne 1982). Implicit in this tradeoff is the belief that greater cognitive effort will lead to accuracy improvements. Consistent with this idea, having a stronger accuracy goal prompts people to think harder (Creyer, Bettman, and Payne 1990). Although much of this research has been conducted in the domain of decisions, there is little reason to believe that a similar lay logic would not apply to judgments as well. In particular, because people believe that greater thought improves accuracy (Creyer et al. 1990; Payne 1982), people may further believe that revising a prior estimate will lead to accuracy gains as a result of the additional thought invested.

Despite the research on the benefits of deliberation and the lay beliefs discussed above, the accuracy benefits of judgment revision have not been systematically explored. On the one hand, it is possible that when revising an estimate, individuals would apply additional cognitive processing in order to improve their initial estimate, especially if they are incentivized to do so. Broadly speaking, the literature reviewed above suggests that such improvement could come from three basic processes: unearthing additional cues

[1] A smaller literature has examined the subset of judgment contexts in which reliance on heuristics (i.e., judgment strategies that require little cognitive effort) improves accuracy. For example, when guessing a value that is a multiplicative function rather than a linear function of cues, interventions which impede greater processing improve judgment quality (Hoffmann, von Helversen, and Rieskamp 2013). Furthermore, instructing people to analyze the reasons underlying their judgments (Ambady 2010; Halberstadt and Levine 1999; Tordesillas and Chaiken 1999; Wilson and Schooler 1991) has been shown to be a type of deliberation that can decrease judgment quality. However, it seems that the consensus of the literature is that additional processing (unconstrained by external instructions dictating the direction of that processing) leads to better outcomes under most circumstances.

from external information (e.g., noticing additional task information or consulting others), retrieving additional cues from memory, or re-combining cues more successfully. On the other hand, there are multiple reasons to hypothesize that simply reconsidering an estimate again, without feedback or additional resources, is unlikely to be fruitful.

First, prior research on the confirmation bias suggest that if an estimator has searched their mind and produced an estimate based on a set of recalled cues, it is unlikely that they would be able to generate disconfirming cues that had gone unnoticed earlier (e.g., Bruner et al. 1956; Koriat et al. 1980; Nickerson 1998). Rather than considering a particular problem from a fresh perspective, individuals tend to revisit prior considerations and generate additional reasons to confirm their earlier conclusions (e.g., Bruner et al. 1956; Koriat et al. 1980; Nickerson 1998). This tendency limits both information search (Nickerson 1998; Skov and Sherman 1986; Snyder and Campbell 1980) and the manner in which cues are combined (Koriat et al. 1980; Lehner, Adelman, Cheikes, Brown 2008; Oswald and Grosjean 2004).

In the domain of quantitative judgments specifically, the extensive literature using the Judge Advisor System (JAS) has shown that people are highly reluctant to revise their earlier judgments even when presented with the diverging estimate of another participant. Indeed, the modal response in most JAS studies is for participants to “stand pat” on their prior estimates and make no adjustment toward advice, a strategy that reliably hurts accuracy (Liberman, Minson, Bryan and Ross 2012; Minson, Liberman and Ross 2011; Soll and Larrick 2009). This behavior may be related to the well-documented tendency toward anchoring and insufficient adjustment (Epley and Gilovich 2006; Frederick and Mochon 2012; Galinsky and Mussweiler 2001; Simmons, LeBoeuf, and Nelson 2010;

Tversky and Kahneman 1974), wherein even irrelevant quantities seem to exert a “pull” on future judgments.

Beyond a tendency toward confirmation bias and anchoring on earlier estimates, individuals may fail to unearth new information or appropriately adjust their estimates because they are overconfident in the accuracy of their original conclusions. The literature on over-precision (Moore and Healy 2008) shows that people are egregiously overconfident in their judgments under uncertainty and that this tendency is extremely difficult to debias (Haran, Moore, and Morewedge 2010; Meikle, Tenney, and Moore 2016). It may be the case that overconfidence either limits individuals’ search for new cues, or makes them reluctant to adjust their judgment when such cues are unearthed, lest adjustment actually leads to an accuracy decrease.

Furthermore, even if additional information was discovered from a second memory search or perusal of external information, how would a judge know whether incorporating this information would improve accuracy or degrade it? Such dilemmas may flummox the judge because people are notoriously unable to distinguish relevant from irrelevant information, and thus often overweight nonpredictive details (Hall, Ariss, and Todorov 2007; Kelly and Simmons 2016; Nisbett, Zukier, and Lemley 1981).

Together, these streams of research suggest several forces that would severely limit any accuracy improvement from revisiting judgments under uncertainty. Yet, these considerations stand in contrast to both practitioner intuition (e.g., Shaw 2019; Slater 2010; Stevenson 2018; Villamor 2020) and prior research suggesting that people might *expect* judgment revision to lead to improvement and act on this expectation. In the current work we directly investigate this apparent tension.

The Present Research

In the present research, we examine both the accuracy gains that individuals achieve when they consider the same estimate twice, as well as observer expectations of those gains. Across 14 studies, we document a robust disconnect between estimator performance and observer expectations. Specifically, we find that although judges do not become more accurate by making an estimate again, observers consistently expect them to. We examine these dual phenomena—judges' failure to improve their accuracy and observer expectations that they do—across a variety of incentive structures, populations, and task instructions. We specifically focus on the potential accuracy gains that may result from *reconsideration* of estimates by isolating reconsideration from other sources of accuracy improvement. In other words, although the provision of external resources often improves accuracy (Arnold, Stenzel, Drolet, and Ramachandran 2016; Kopelman 1986; Levine, Samet, and Brahlek 1975; Peterson and Pitz 1986; Remus, O'Connor, and Griggs 1996), the current research investigates a different question—whether accuracy can be improved by exclusively "trying again."[2]

The disconnect between judge performance and observer expectations that we document in our studies has important consequences. If managers incorrectly believe that others can increase the accuracy of their judgments by simply thinking about them again, they may invest employee time into a task that is essentially useless. Furthermore, they may fail to provide employees with the resources and tools that would *actually* make

[2] In the General Discussion, we discuss contexts in which the combination of reconsideration in conjunction with the guidance of external resources improves judgment accuracy. Before doing so, we empirically examine whether reconsideration alone (absent other interventions such as the provision of external resources) improves accuracy.

them more accurate, expecting that simply thinking again will do the trick. In the absence of appropriate resources and tools, such misguided expectations of accuracy gains may lead to inaccurate forecasts, negative performance evaluations, and suboptimal managerial decisions.

Beyond the applied implications, our studies offer basic insight into the cognitive processes underlying second estimates that have thus far received limited attention. The literature on the "crowd within" has proposed descriptive models which characterize individuals as drawing second estimates seemingly at random from a distribution in their minds (Gaertig and Simmons 2020). By contrast, we find that lay people believe second estimates are guided by an entire suite of deliberate cognitive processes that systematically improve the accuracy of second estimates. By clearly delineating the gap between estimator performance and observer expectations, we can begin to understand the underlying psychology of trying again in the domain of quantitative judgment.

Therefore, after first documenting that reconsidering estimates does not improve their accuracy, we next provide the first empirical evidence that observers falsely believe that it does and uncover the negative consequences of this misprediction. The current research is also the first to unearth the specific channels through which observers believe that the greater thought resulting from reconsideration translates into accuracy gains in the context of judgment uncertainty.

Overview of Studies

We present 14 studies examining the effect of making a second estimate on both judgment accuracy and observer expectations. Studies 1–2 (as well as Supplementary Studies A–C) examine the estimates of both lay people (Study 1A and Supplementary

Studies A–B) and experts (Study 2A), and find that revisiting an estimate does not seem to lead to greater accuracy. Nevertheless, both lay people (Study 1B, Study 2B, and Supplementary Study C) as well as experts (i.e., judgment and decision-making researchers; Study 1C) seem to expect that second estimates will be more accurate than first estimates. This misprediction persists when observers are financially incentivized for correctly gauging relative estimation accuracy (Study 2B and Supplementary Study C).

Study 3 captures observers' open-ended responses describing why they believe second estimates to be more accurate. When we systematically manipulate these beliefs in Studies 4A–4C (by telling observers that estimators do not engage in the cognitive processes believed by Study 3 participants to drive improvement), expectations of increased accuracy attenuate. Notably, Study 5 finds that after making two estimates, people believe that their own second estimate is actually inferior to the first. Study 5 further finds that this belief can be leveraged to debias the exaggerated expectations of others' accuracy improvement. Finally, Study 6 documents the interpersonal costs of the mismatch between individuals' beliefs about the benefits of second estimates for themselves versus other people.

Open Science Statement

We pre-registered Study 1A, Study 1B, Study 2A, Study 2B, Study 3, Study 5, Study 6, and Supplementary Study A before beginning data collection. Our pre-registration materials, surveys, data, and analysis scripts are posted on Researchbox.org (https://researchbox.org/46&PEER_REVIEW_passcode=MOVCVU).

STUDY 1: SECOND ESTIMATES ARE NOT MORE ACCURATE, BUT OBSERVERS BELIEVE THEY ARE

In Study 1A we tested whether revising a quantitative estimate leads to accuracy gains. We varied the way we prompted participants to revise their estimates by telling them to either think again, think harder, or offering them an incentive for greater accuracy. We further explored the generalizability of our findings across five familiar estimation domains.

Method

We aimed to recruit 300 participants from Mechanical Turk. Participants who failed an attention check at the beginning of the survey were barred from entering the survey, and we collected data from a total of 301 participants (44.5% female, average age = 36.7 years). Participants made quantitative estimates regarding five different topics: The number of calories in a pictured meal, the price of a pictured vehicle, and the annual average costs of electricity, food and childcare for an American family. Participants did not receive accuracy incentives for making these initial estimates, thus setting a fairly low bar from which to improve. After participants entered their estimates into an empty survey field, we asked them to make a second estimate for each of the five items.

Several methodological features were common to all our studies in which people made estimates. First and foremost, the request to make a second estimate was unexpected. In addition, participants did not receive any feedback about the accuracy of their first estimate before making their second estimate, mirroring many real-world situations in which accuracy feedback is unavailable. Finally, participants completed the first and second estimate during the same survey session and were able to see their first estimate on the same screen on which they entered their second estimate.

In Study 1A, we randomly assigned participants to one of three conditions which differed only in the instructions prompting them to make a second estimate. In the *Guess Again* condition, participants were simply asked to make a second estimate. In the *Try Harder* condition, participants were asked to make a second estimate *and* to try harder. In the *Incentivized* condition, participants were asked to make a second estimate *and* to try harder *and* were offered an added five-cent bonus if their second estimate was more accurate than the first. In all conditions, participants were asked to enter a second estimate that differed from their first estimate.[3] Participants then viewed the information that they had previously viewed regarding each of the five estimation items (as well as their first estimate for each item), and entered their second estimates into empty survey fields. At the end of this study and all of the studies in the manuscript, participants reported their gender and age.

Results

We followed our pre-registered analysis plan to exclude estimate pairs (i.e., a participant's first and second estimate for a particular estimation item) that differed from each other by an order of magnitude (i.e., by a multiple of ten or more), because we anticipated that such large deviations reflected typographical errors. This exclusion rule resulted in the exclusion of 98 out of the 3,010 estimates (i.e., 3.3% of the estimates). To

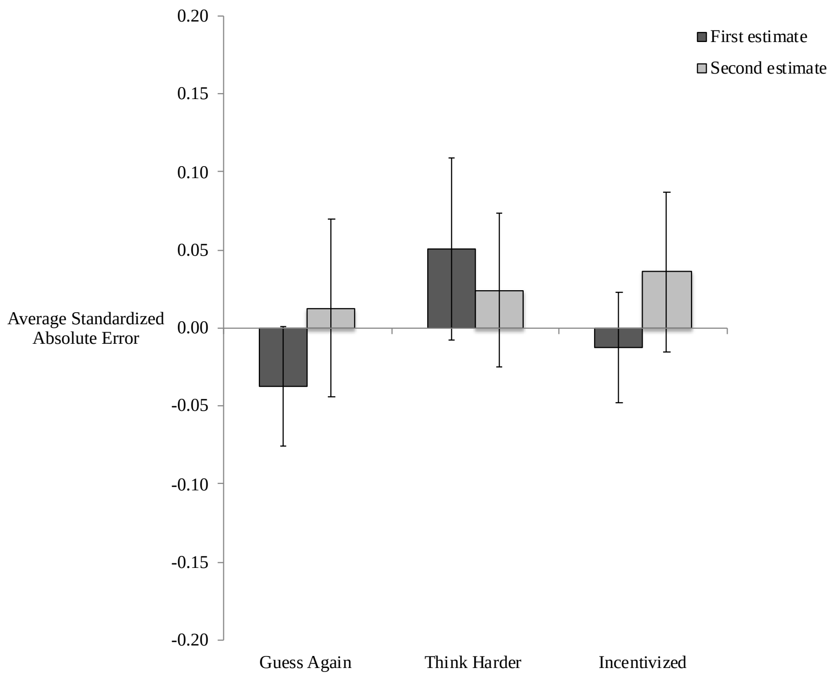
[3] We included this instruction in order to ensure that any lack of improvement was not due to participants simply entering the same estimate for both their first and second estimates in order to conserve effort. When the instruction to enter a different second estimate was removed we find similar results (Supplementary Study A).

analyze estimation accuracy, we followed our pre-registered analysis plan by first calculating the absolute difference between each participant's estimate for each item and the corresponding correct answer for that item. Because the correct answers to each item were on vastly different scales, we standardized the errors in the following manner. We first z-scored the absolute errors within estimation item across conditions for the first estimation round. We then used the mean and standard deviation of the absolute errors for each item on the first estimation round to standardize the absolute errors on the second estimation round. The goal of this procedure was to us to test whether accuracy improved from the first round of estimation to the second one, and whether this improvement differed as a function of condition. This procedure resulted in first-round estimates with average error (for any given estimation item collapsed across conditions) having a standardized error of zero. Second round estimates that were more accurate had negative standardized errors, and second round estimates that were less accurate had positive standardized errors.

In order to examine whether making a second estimate reduced error, we regressed the standardized errors on estimation round (first estimates = 0; second estimates = 1), clustered the data at the level of participant, and included fixed effects for each of the five estimation items. We found no significant decrease in error from the first to the second round of estimation in any condition (Figure 1): We observed no decrease in error when we prompted participants to think again (b = .05, SE = .06, t = .84, p = .40; 95% CI: –.07 to .17), prompted participants to think again and think harder about their estimates (b = –.03, SE = .04, t = –.75, p = .46; 95% CI: –.10 to .04), or incentivized participants for greater accuracy (b = .05, SE = .05, t = .92, p = .36; 95% CI: –.06 to .15).

When we collapsed the data across the three conditions, error again did not decrease (b = .02, SE = .03, t = .82, p = .41; 95% CI: –.03 to .08).

Figure 1. Descriptive statistics as a function of condition and estimation round in Study 1A.



Note. The height of the bars represents the average standardized absolute estimation error by condition and estimation round. The error bars are unclustered standard errors.

Study 1B

Study 1A finds that second estimates are no more accurate than first estimates across three different elicitation procedures and five different estimation items. In Study 1B we turn to examining observer expectations. To that end, we described Study 1A to a separate sample of Mechanical Turk workers, who made incentivized guesses about the relative accuracy of their fellow workers' first and second estimates.

Method

We aimed to recruit 300 participants from Mechanical Turk. Participants who failed an attention check at the beginning of the survey were barred from entering the survey, and we collected data from a total of 320 participants (59.4% female, average age = 36.0 years). All participants read about Study 1A. In particular, they read that other Mechanical Turk workers had made numerical estimates of five different quantities and that they now would be asked to assess the accuracy of those estimates.

Next, participants were randomly assigned to one of three conditions, which differed only in whether participants saw that the earlier participants had read the instructions in the *Guess Again, Try Harder,* or *Incentivized* condition (i.e., the instructions that differed by condition in Study 1A). We presented participants with the five estimation items, as well as the average absolute error of the first estimates produced by the earlier participants. For example, participants in the *Guess Again* condition learned that the average percent error of the prior MTurkers' first estimate of the price of the pictured vehicle was (on average) 85% of the correct answer before these prior participants were unexpectedly asked to make a second estimate.

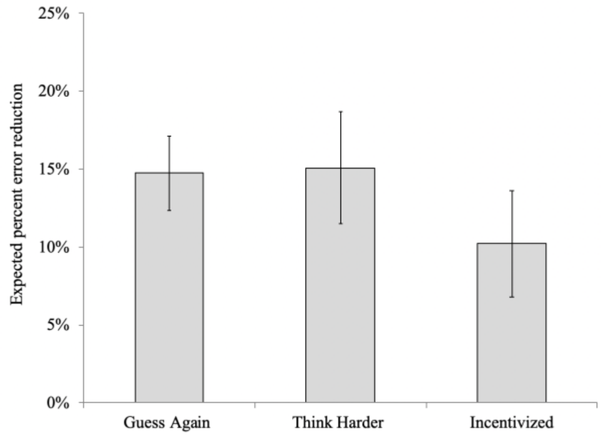
We then asked participants to estimate the average amount of error of these prior participants' second estimates for each of the five items. Study 1B's participants learned that if the accuracy of their estimates was in the top 5% of survey takers, they would earn a $5 bonus. Participants entered their predicted error for each item into an empty field.

Results

We followed our pre-registered analysis plan to exclude estimates that differed by three or more standard deviations from the mean estimate for the corresponding estimation item, attributing those estimates to inattention or typographical errors. This exclusion rule resulted in the exclusion of 25 out of the 1,575 estimates (i.e., 1.6% of the estimates).

In order to examine whether participants expected that revisiting an estimate would lead to error reduction, we calculated the average expected change in error for participants overall, and by item and elicitation condition. We then clustered the standard errors at the level of participant, and conducted a regression to examine whether the expected change in error differed from zero. Across all conditions, participants expected that others' second estimates would have less error than their first estimates (Figure 2): Participants expected others' error to decrease when others were prompted to think again (b = 14.50, SE = 2.37, t = 6.11, p < .001; 95% CI: 9.79 to 19.20), prompted to think again and think harder about their estimates (b = 15.53, SE = 3.44, t = 4.52, p < .001; 95% CI: 8.72 to 22.34), or offered incentives for greater accuracy (b = 10.80, SE = 3.04, t = 3.55, p < .001; 95% CI: 4.76 to 16.85). Participants expected others to decrease the magnitude of their error in their second (vs. first) estimates by the same amount across conditions (bs < 4.73, ts < 1.03, ps > .30). Not surprisingly, the result persisted when we collapsed the data across conditions (b = 13.81, SE = 1.75, t = 7.91, p < .001; 95% CI: 10.38 to 17.24).

Figure 2. Average expected percent error reduction in Study 1B.



Note. This figure charts raw means. The error bars represent unclustered standard errors.

Study 1C

It is possible that the misprediction documented in Study 1B is limited to lay people who fail to appreciate the challenges involved in making quantitative judgments under uncertainty. Study 1C examines whether individuals with expertise in the field of judgment and decision making also expect that others' second estimates are more accurate than first estimates. We described the procedures that participants completed during Study 1A to researchers attending the annual Society for Judgment and Decision Making conference (the flagship conference for this academic field), and asked them about the relative accuracy of participants' first and second estimates.

Method

Two hundred and twelve attendees at the annual Society for Judgment and Decision Making conference (50.0% female, average age = 51.4 years) completed a survey on hand-held tablets during conference breaks. The survey was administered by a research assistant who was blind to the hypothesis.[4]

[4] At the conclusion of the survey, participants indicated their current professional position by selecting one of five buttons labeled with "faculty member," "post-doc," "doctoral student," "university student," or

Participants read that Mechanical Turk workers were asked to estimate five different quantities and that they were then asked to make second estimates of those same quantities. We then showed all participants a summary of the instructions employed to elicit these second estimates in all three of the conditions in Study 1A. Next, participants viewed each of the estimation items from Study 1A and also saw the average percent error of the workers' first estimates from a pilot of that study (which used the same items and produced the same result as the pre-registered Study 1A).

Our dependent variable was participants' evaluations of the relative accuracy of first versus second estimates in Study 1A. Participants entered their responses on separate 7-point scales (*1: These second estimates were a lot less accurate; 7: These second estimates were a lot more accurate*). Because of time limitations, all participants completed this perceived accuracy measure for three estimation items out of the five. After the third and the fourth items, we gave participants the choice of whether to exit the survey flow and proceed to the demographic information, or to complete a prediction about one more item. Thus, participants who completed the entire task provided 15 accuracy ratings (i.e., they evaluated the accuracy of the workers' estimates for each of the five items in all three conditions). On average, participants completed 11.67 accuracy ratings (SD = 2.30).

Results and Discussion

We recoded participants' accuracy ratings (initially captured on a 1–7 scale) onto a –3 to +3 scale. If participants believed that second estimates were more accurate than

"other." Descriptive statistics revealed that 21.3% of participants were faculty members, 8.5% were post-doctoral researchers, 50.7% were doctoral students, 5.2% were undergraduates, and 14.2% identified as "other" (e.g., industry professionals and research assistants).

first estimates, their ratings would be higher than the scale midpoint of zero. Table 1 provides the mean relative accuracy ratings offered by judgment and decision-making researchers, by condition. To test whether these ratings were significantly different from the midpoint of zero, we ran an empty regression for each condition separately, clustering the data at the level of participant. In each condition, the intercept was positive and significantly differed from the midpoint of zero (*Guess Again* condition: $b = .41$, $se = .07$, $z = 6.13$, $p < .001$, 95% CI: .28 to .54; *Try Harder* condition: $b = .41$, $se = .06$, $z = 6.56$, $p < .001$, 95% CI: .29 to .53; *Incentivized* condition: $b = .67$, $se = .07$, $z = 10.15$, $p < .001$, 95% CI: .54 to .81). When we collapsed the data across the three conditions, the expectation that accuracy would improve on the second estimate persisted ($b = .50$, $SE = .05$, $t = 9.28$, $p < .001$, 95% CI: .39 to .60).

An examination of Table 1 also suggests that researchers expected that incentives would lead to greater accuracy improvements than merely thinking again or thinking harder, with the results from the two latter conditions appearing nearly identical. To test the significance of this pattern we created dummy codes for the *Try Harder* condition and the *Guess Again* condition (both coded as zero), and then compared the forecasted accuracy improvement of estimates in those conditions to the forecasted accuracy improvement in the *Incentivized* condition (coded as one). We again clustered the data at the level of participant. This analysis confirmed that researchers did indeed expect greater accuracy improvements when the Mechanical Turk workers were incentivized to produce more accurate second estimates than when they were merely asked to think again ($b = .26$, $se = .07$, $z = 3.74$, $p < .001$, 95% CI: .12 to .40) or try harder ($b = .27$, $se = .06$, $z = 4.56$, $p < .001$, 95% CI: .15 to .38).

Table 1. Average expected relative accuracy of second estimates by elicitation condition and estimation item in Study 1C.

	Elicitation Condition		
	Guess Again	Try Harder	Incentivized
Estimation Item			
Calories	0.39	0.39	0.73
Car cost	0.38	0.42	0.66
Childcare	0.54	0.43	0.70
Electricity	0.39	0.47	0.68
Food	0.42	0.37	0.61

STUDY 2: EXPERTS ALSO DO NOT GENERATE MORE ACCURATE SECOND ESTIMATES, BUT OBSERVERS BELIEVE THEY DO

In Study 1A people failed to improve the accuracy of their estimates when asked to reconsider their first estimates across a variety of estimation tasks and elicitation instructions. Importantly, however, the estimators in Study 1A were lay people. Is it possible that experts succeed where amateurs fail? For example, perhaps experts have learned from experience that estimates can be improved by considering the problem from a different vantage point. Furthermore, experts may possess more knowledge or greater skill at separating the relevant from irrelevant information when they reconsider their judgments. We investigate this possibility in Study 2A.

We also examine whether it is possible that the pattern of results observed in Study 1A is caused by a simple lack of motivation. Perhaps Study 1A estimators would improve their accuracy if the incentives for doing so were greater. Thus, in Study 2A, we offered estimators a larger bonus for accuracy.

Importantly, we also asked participants in Study 2A, who were all professional realtors, to make estimates in a more information-rich context which was also the domain of their professional expertise: estimating home prices based on real estate listings. This estimation context enabled us to investigate whether the current findings persist among experts engaged in relatively complex estimation tasks, which may increase their likelihood of noticing or recalling information that was not reflected in their first estimates.

Study 2A

Method

We aimed to recruit 100 real estate agents working in Seattle, and collected data from a total of 105 agents (58.1% female, average age = 47.6 years). The study was conducted in April of 2020, and the real estate agents learned that they would be estimating the sale price of Seattle homes sold in January of 2020. We told participants that the agent who made the most accurate estimates would receive a $100 Amazon gift card.

Participants next viewed online listings for five homes downloaded from Redfin.com. Participants saw pictures of the homes' interior and exterior, the neighborhood in which each home was located, and a written description of each home's attributes. We redacted the address of each home, its listing price, and its Multiple Listing Service number so that participants could not simply look up the correct answer.

Participants estimated the January 2020 sale price of each of the five homes and entered their estimates into an empty field. After participants entered their estimates, they learned that we were interested in how people make second estimates. They were then (unexpectedly) asked to make a second estimate for each of the five homes that they had previously viewed. In order to isolate the unique effect of reconsideration (absent influence from additional resources), participants were asked to not reference resources they had not utilized during their first estimation round. We incentivized participants to achieve greater accuracy during this second estimation round by offering them a $100 Amazon gift card if their second estimates were more accurate on average than other participants' second estimates.

Results

As in Study 1A, we excluded pairs of estimates from our analyses that were an order of magnitude larger or smaller than each other (assuming that this deviation resulted from participants entering an incorrect number of zeros). This resulted in the exclusion of 12 out of the 1,050 estimates (i.e., 1.14% of estimates).[5]

We again examined whether making a second estimate reduces error by regressing the standardized estimation error on estimation round (first estimates = 0; second estimates = 1), clustering the data at the level of participant and including fixed effects for each estimation item. We found no significant decrease in error from the first estimate (where we z-scored the errors, resulting in a mean of zero) to the second estimate (B = –.02, SE = .03, t = –.53, p = .59; 95% CI: –.08 to .05).

Further, exploratory analyses revealed that 40% of first estimates were more accurate than second estimates, 40% of second estimates were more accurate than first estimates, and 20% of first estimates and second estimates were identical (and thus

[5] We employed this exclusion criterion to remain consistent with Study 1A. However, we pre-registered a different exclusion criterion in Study 2A (i.e., the exclusion of estimates that were an order of magnitude larger or smaller than the correct answer)—and employing this exclusion rule again found no significant decrease in error from the first estimate to the second estimate (B = –.00, SE = .04, t = –.04, p = .97; 95% CI: –.07 to .07).

equally accurate).[6] These latter results suggest that the observed lack of improvement is not due to a minority of extreme declines in estimation accuracy occluding a majority of estimates that modestly improved. Rather, it appears that making a second estimate simply produces no systematic accuracy improvement for professional real estate agents, just as it produces no systematic accuracy improvement for Mechanical Turk participants.

Study 2B

Study 2A provides converging evidence that the additional thought devoted to making a second estimate does not improve its accuracy. Study 2B examines whether observers who are incentivized for correct assessments of relative accuracy expect that experts' second estimates are more accurate than those experts' first estimates.

Method

We aimed to recruit 400 participants from Mechanical Turk. Participants who failed an attention check at the beginning of the survey were barred from entering the survey, and we collected data from a total of 407 participants. We followed our pre-registered analysis plan to exclude participants who failed either one of two attention checks, which resulted in a final sample of 316 participants (48.9% female, average age = 37.2 years).

[6] The same exploratory analyses in the additional studies in which participants made first and second estimates similarly found that making a second estimate did not systematically increase accuracy: In Study 1A, 41% of second estimates were less accurate than first estimates, 44% of second estimates were more accurate than first estimates, and 14% of second estimates were equally as accurate as first estimates; in Study 3A, 43% of second estimates were less accurate than first estimates, 44% of second estimates were more accurate than first estimates, and 13% of second estimates were equally as accurate as first estimates; in Supplementary Study A, 33% of second estimates were less accurate than first estimates, 39% of second estimates were more accurate than first estimates, and 28% of second estimates were equally as accurate as first estimates.

All participants read about Study 2A. They next learned that one of the real estate agents who participated in Study 2A had been randomly assigned to be their partner, and they had the opportunity to receive a bonus based on the accuracy of their partner's estimates. Participants then learned that they could choose whether to base this bonus on their partner's first estimate or second estimate, and that they would receive an additional twenty-cents if their chosen estimate was within 10% of the home's actual sale price. Thus, participants were directly incentivized to choose the most accurate estimate. Participants entered their choice by selecting a button labeled either "My partner's first estimate" or "My partner's second estimate."

Results

Participants were more likely to choose to base their bonus on the experts' second estimates (57.6%) than on the experts' first estimates (42.4%), $\chi^2(df = 1, N = 316) = 7.29$, $p = .007$ (Cohen's $d = .31$, 95% CI: .08 to .53). As in the earlier studies, the majority of participants anticipated that expert estimators would improve their accuracy when making an estimate again.

Discussion of Studies 1–2

Our studies thus far provide converging evidence that estimation accuracy does not improve as a result of revisiting prior judgments, irrespective of participant population, method of eliciting second estimates, or estimation topic. Yet, both lay people as well as judgment and decision-making researchers expect others to be able to attain such improvement. In a final study in this series (Supplementary Study B), we examined whether offering a very large accuracy bonus would improve the relative accuracy of second estimates. We again recruited participants on Mechanical Turk to make estimates, this time regarding the behaviors and preferences of their fellow workers. We incentivized the accuracy of second estimates with $5, a 4,900% increase over the 10 cents we offered for making accurate first estimates. Notably, the estimation domain (behaviors and preferences of Mechanical Turk workers in a prior study), ensured that even highly-incentivized participants could not look up the answers. In spite of this, and conceptually replicating Studies 1–2, incentivized observers again incorrectly expected that these estimators' second estimates would in fact be more accurate than first estimates (Supplementary Study C).

STUDY 3: WHY DO OBSERVERS EXPECT RECONSIDERATION TO IMPROVE ESTIMATION ACCURACY?

In Studies 3 and 4 we examine the causes behind observers' inflated expectations of accuracy improvement. Prior literature suggests that accuracy improvement in the absence of feedback can arise from one of three basic processes. First, it is possible for people to simply notice more details about the information provided to them and incorporate those details into their judgments (Horstmann, Ahlgrimm, and Glöckner 2009; Levin, Huneke, and Jasper 2000). Second, it also possible for individuals to recall additional pertinent information beyond that which is provided as part of the task and use that information to improve their estimates (e.g., Henkel 2004; Taber 2011). Finally, holding constant the cues under consideration, individuals may put different weights on various cues, leading to improvement (e.g., Harvey, Harries, and Fischer 2000; See, Morrison, Rothman, and Soll 2011).

In Study 3, we investigate whether these possibilities underlie observers' intuitions. To that end, we described a judgment task to participants and simply asked them to report their intuitions about why first or second estimates may be more accurate, in order to explore the extent to which these intuitions conform with the suggestions of prior research. After participants reported their intuitions in an open-ended format, we asked them to code their responses into the three categories above, as well as an additional "other" category. We were interested in whether the three categories together offer a comprehensive typology of why people believe second estimates are more accurate than first estimates. Furthermore, we wanted to know whether one of the categories is favored by participants over the others. In Study 4, we manipulate the reasons generated by Study 3 participants in order to determine which of these intuitions actually *cause* observers to believe that reconsideration improves second estimates.

Method

We aimed to recruit 200 participants from Mechanical Turk. Participants who failed either one of two attention checks at the beginning of the survey were barred from entering the survey, and we collected data from a total of 217 participants. Per our pre-registration plan, we removed data from 23 participants who wrote fewer than ten words during the task, resulting in a final sample of 194 participants (mean age = 37; 54% female).

All participants read about a study wherein another Mechanical Turk worker had been incentivized with five cents to correctly estimate the listing price of a home, and that the worker was asked to make a second estimate of the price after entering their first estimate. Participants then reported which estimate they thought was most accurate. Participants entered their responses by selecting either a button labeled "Their first estimate" or "Their second estimate."

Next, we asked participants to describe why they thought that the estimate they selected was the more accurate of the two. We offered participants three open response boxes to describe their reasons and asked them to enter only one reason per box. These responses were copied onto the subsequent screen, where we presented participants with broad categories of reasons why one estimate may be more accurate than another. We asked participants to select a category which best encompassed each reason that they had generated.

The categories differed as a function of whether a participant had chosen the first or second estimate as being more accurate. We focus our analysis on the majority of participants who expected second estimates to be more accurate than first estimates. We report the results from the remaining participants in the Supplementary Materials.

Participants who indicated that the second estimate was more accurate chose between the following four categories into which to classify their stated reasons: “When people make their second estimate, they may notice new details in the information that they did not previously notice when they had made their first estimate”; “When people make their second estimate, they may remember relevant information that did not come to mind when they had made their first estimate”; “When people make their second estimate, they may more effectively weigh the information that they had already considered when they had made their first estimate”; “Another reason not listed above.” Participants who chose the last option were provided with an empty field to describe a category that captured their stated reason. The options presented to participants who indicated that the first estimate was more accurate are detailed in the Supplementary Materials.

Results

Consistent with prior results, over twice as many participants (72.2%) believed that second estimates were more accurate than first estimates, than believed otherwise, $\chi^2(df = 1, N = 194) = 38.12$, $p < .001$ (Cohen's $d = .99$, 95% CI: .67 to 1.30). These 140 participants generated 207 reasons to explain their evaluation.

Specifically, participants categorized more than half (53.6%) of those reasons as consistent with the idea that estimates are improved by noticing new information during the second estimate. Participants categorized 16.9% of their reasons as consistent with estimates being improved by new information remembered during the second estimate. Finally, participants categorized 25.1% of the reasons as consistent with the idea that estimates are improved by more effective weighting of information. Only 4.3% of reasons were categorized into the "other" category.

STUDY 4: DEBIASING INFLATED EXPECTATIONS OF ACCURACY IMPROVEMENT

Study 3 illuminates people's self-reported beliefs about why others' estimates become more accurate upon reconsideration. These responses predominately fell into three categories describing cognitive processes that observers expected estimators to engage in. In Study 4, we provide observers with information about whether or not estimators engaged in these three categories of cognitive processes in order to test whether these beliefs *cause* their misguided expectation of accuracy improvement. Specifically, we manipulate whether observers learn that when making a second estimate, an estimator noticed new information (Study 4A), recalled previously-unrecollected information (Study 4B), or weighted the available information differently (Study 4C). If people believe that these cognitive processes do in fact cause estimates to become more accurate, then informing them that estimators did not engage in these processes should dampen their accuracy improvement expectations.

It is also possible that observers in our prior studies believed that estimates improve with revision simply because estimators devote little effort to their first estimates. Thus, Studies 4A–4C manipulate information about the relevant cognitive processes on both estimation rounds to test whether perceivers still expect improvement on the second round even when first round estimates already feature the types of processing that observers believe to maximize accuracy.

Method

Five hundred ninety seven Mechanical Turk participants (61.5% female, average age = 39.3 years) completed Study 4A; 401 Mechanical Turk participants (66.7% female, average age = 39.1 years) completed Study 4B; and 402 Mechanical Turk participants (57.0% female, average age = 37.4 years) completed Study 4C. Participants who failed any one of three attention checks at the beginning of each study were barred from entering the study.

In all three studies, we randomly assigned each participant to view one of five house listings from RedFin.com. Participants read that another Mechanical Turk worker had estimated the listing price of this house, after which the worker was instructed to make a second estimate.

In each study, we further randomly assigned participants to one of four conditions that differed with regard to which estimation round they received additional information about. In the *Baseline* condition, participants were simply told that another Mechanical Turk worker estimated the listing price of the pictured house, after which the worker was instructed to make a second estimate. This condition was identical across Studies 4A–4C. In the *First Round* condition, participants were given additional information about the estimation process that the estimator implemented during the first round of estimation. In the *Second Round* condition, participants were given additional information about the estimation process that the estimator implemented during the second round of estimation. Finally, participants in the *Both Rounds* condition were given additional information about both rounds of estimation. Thus, participants were randomly assigned to one cell in a 2 (Additional information about first round: Present vs. Absent) × 2 (Additional information about second round: Present vs. Absent) between-participant design.

In Study 4A, the additional information described whether or not the estimator had noticed new relevant details regarding the target house. In Study 4B, the additional information described whether or not the estimator recalled new relevant information. In Study 4C, the additional information described whether or not the estimator had differentially weighed the relevant judgment cues in developing their estimate. The text of the manipulation in each condition of Studies 4A–4C is presented in Table 2.

In all studies, after viewing this information, participants indicated which of the two estimates they thought was more accurate. Participants entered their responses on a 7-point scale (*1: Their first estimate is a lot more accurate; 4: Both estimates are equally accurate; 7: Their second estimate is a lot more accurate*).

Table 2. Manipulated instructions in Studies 4A–4C.

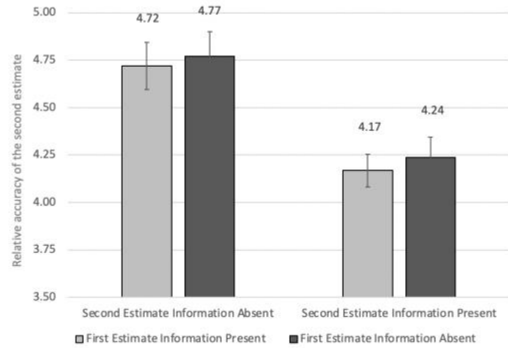
	Condition			
Study	*Baseline* condition	*First Round* condition	*Second Round* condition	*Both Rounds* condition

Study 4A	The MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view this information about this house a second time and make a second estimate about this house's listing price.	Before the MTurker made this estimate, they looked carefully at the information above for several minutes, and noticed many of the details in the information above. After doing so, the MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view this information about this house a second time and make a second estimate about this house's listing price.	The MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view this information about this house a second time and make a second estimate about this house's listing price. When the MTurker looked at the information again, the MTurker did not notice any new details that they hadn't previously noticed when they looked at the information earlier.	Before the MTurker made this estimate, they looked carefully at the information above for several minutes, and noticed many of the details in the information above. After doing so, the MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view this information about this house a second time and make a second estimate about this house's listing price. When the MTurker looked at the information again, the MTurker did not notice any new details that they hadn't previously noticed when they looked at the information earlier.
Study 4B	The MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view this information about this house a second time and make a second estimate about this house's listing price.	Before the MTurker made this estimate, they thought carefully about how houses' listing prices are determined for several minutes, and recalled many details about how houses' listing prices are determined. After doing so, the MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view the information about this house a second time and make a second estimate about this house's listing price.	The MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view the information about this house a second time and make a second estimate about this house's listing price. When the MTurker estimated this house’s listing price again, they did not recall any new details about how houses' listing prices are determined that they hadn't already recalled when they made their first estimate.	Before the MTurker made this estimate, they thought carefully about how houses' listing prices are determined for several minutes, and recalled many details about how houses' listing prices are determined. After doing so, the MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view the information about this house a second time and make a second estimate about this house's listing price. When the MTurker estimated this house’s listing price again, they did not recall any new details about how houses' listing prices are determined that they hadn't already recalled when they made their first estimate.
Study 4C	The MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view this information about this house a second time and make a second estimate about this house's listing price.	Before the MTurker made this estimate, they thought about the right way to approach making this estimate and came up with a step-by-step process for estimating this house's listing price. After doing so, the MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view this information about this house a second time and make a second estimate about this house's listing price.	The MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view the information about this house a second time and make a second estimate about this house's listing price. Therefore, the MTurker looked at the information again. However, this time the MTurker decided to follow their intuitive judgment (i.e. their “gut feeling”) to estimate this house's listing price.	Before the MTurker made this estimate, they thought about the right way to approach making this estimate and came up with a step-by-step process for estimating this house's listing price. After doing so, the MTurker entered their estimate of this house's listing price into an empty field. Next, like all of the other MTurkers in the survey, the MTurker was asked to view the information about this house a second time and make a second estimate about this house's listing price. Therefore, the MTurker looked at the information again. However, this time the MTurker decided to follow their intuitive judgment (i.e. their “gut feeling”) to estimate this house's listing price.

Results

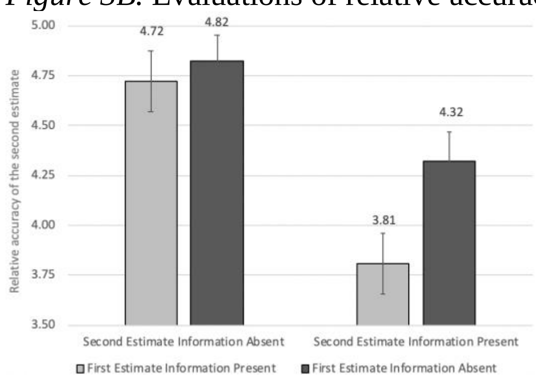
Study 4A. Figure 3A presents the means and standard errors by condition. The figure makes it clear that exaggerated accuracy improvement expectations were attenuated among participants who believed that an estimator did not notice additional details during the second estimation round. In order to examine the statistical significance of this pattern, we conducted a 2 (Additional information about first round: Present vs. Absent) × 2 (Additional information about second round: Present vs. Absent) ANOVA on the expected accuracy data. This analysis revealed no significant interaction ($F(1, 593) = .01$, $p = .92$), but a significant main effect of the second round information emerged: When participants learned that the estimator had not noticed new details during the second estimation round, their expectations of accuracy improvement were attenuated, $F(1, 593) = 24.83$, $p < .001$ (Cohen’s $d = .41$, 95% CI: .25 to .57). The analysis revealed no main effect of the first round information: Learning that the estimator had noticed many details while generating their first estimate did not attenuate accuracy improvement expectations, $F(1, 593) = .29$, $p = .589$.

Figure 3A. Evaluations of relative accuracy (Study 4A).



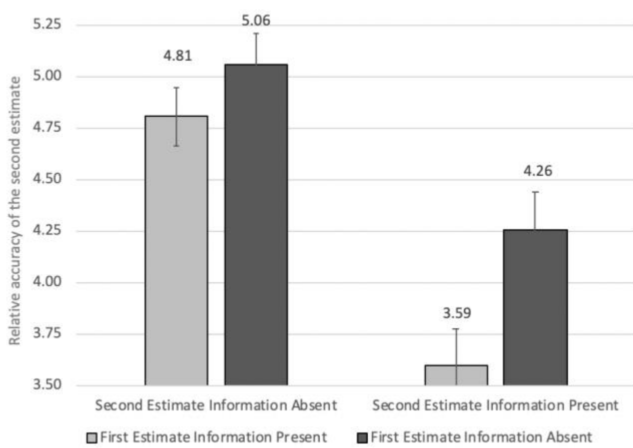
Study 4B. The same analysis on the Study 4B data again did not reveal a significant interaction ($F(1, 397) = 1.99$, $p = .16$), but a significant main effect of second round information emerged: Accuracy improvement expectations attenuated when participants believed that an estimator did not recall additional information during the second estimation round, $F(1, 397) = 23.40$, $p < .001$ (Cohen's $d = .47$, 95% CI: .27 to .66; Figure 3B). A main effect of additional first round information also emerged: Learning that the estimator had recalled relevant information while formulating their first estimate reduced accuracy improvement expectations, $F(1, 397) = 4.41$, $p = .04$ (Cohen's $d = .19$, 95% CI: .00 to .39).

Figure 3B. Evaluations of relative accuracy (Study 4B).



Study 4C. The same analysis on the Study 4C data again did not reveal a significant interaction ($F(1, 398) = 1.54$, $p = .22$), but again revealed a significant main effect of second round information: When participants learned that the estimator did not more systematically weigh the judgment cues during the second estimation round, their expectations of accuracy improvement attenuated, $F(1, 398) = 36.94$, $p < .001$ (Cohen's $d = .61$, 95% CI: .41 to .81; Figure 3C). A main effect of first round information also emerged: When participants learned that the estimator systematically weighed the judgment cues while formulating their first estimate, expectations of accuracy improvement also attenuated, $F(1, 398) = 7.58$, $p = .006$ (Cohen's $d = .26$, 95% CI: .06 to .46).

Figure 3C. Descriptive statistics in Study 4C.



Discussion

Study 4 offers experimental evidence supporting the open-ended responses gathered in Study 3. Observers expect accuracy improvement from second estimates because they believe that estimators may engage in three types of cognitive processing that benefit accuracy. When we manipulate these beliefs by telling observers that an estimator did not notice additional details (Study 4A), did not recall new relevant information (Study 4B), and did not systematically reweight the available cues (Study 4C), they expect second estimates to confer diminished accuracy benefits. Importantly, these studies also suggest that accuracy improvement expectations do not exclusively stem from observers believing that estimators devote little effort during the first estimation round.

Although we are reluctant to make comparisons across studies, it is interesting that the effect of providing second round information seems to be the largest in Study 4C, where that information addresses whether another individual effectively weighted available cues during the two estimation rounds. The fact that observers expect that weighting cues to be key to accuracy improvement stands in stark contrast to prior research suggesting that estimators are actually quite poor at appropriately weighting cues (Kelly and Simmons 2016).

In sum, Study 3 and Study 4 provide converging descriptive and experimental evidence that people believe revising estimates improves accuracy because estimators engage in specific beneficial cognitive activities when doing so. However, our estimation data suggest that these benefits are not realized in practice.

STUDY 5: DEBIASING BY LEVERAGING ESTIMATORS' BELIEFS ABOUT THEIR OWN ACCURACY IMPROVEMENT

Study 5 tests an additional intervention to debias the misprediction documented in earlier studies. Given that people do not improve in the course of making a second estimate, we test whether having the experience of actually making a second estimate leads individuals to realize that others will similarly fail to improve.

It is possible, that in assessing their own first estimates, individuals believe themselves to have already incorporated all relevant cues and weighted them optimally, given the information available to them. After all, if there were extra considerations to be incorporated, would they not have made a different first estimate? The large literature on overconfidence in first judgments supports this possibility (Haran et al. 2010; Meikle et al. 2016; Moore and Healy 2008; Tenney, Meikle, Hunsaker, Moore, and Anderson 2019).

If people do believe that their own first estimate is already optimal, they may conclude that any change has a higher likelihood of decreasing accuracy rather than increasing it. Because self-other biases can attenuate when individuals judge others' experiences after having the same experience themselves (Faro and Rottenstreich 2006; Li, Rohde, and Wakker 2017), Study 5 tests whether the experience of making initial and revised estimates can debias inflated expectations regarding *others'* accuracy improvement.

Method

We aimed to recruit 600 participants from Mechanical Turk. Participants who failed an attention check at the beginning of the survey were barred from entering the survey, and we collected data from 605 participants (52.7% female, average age = 38.0 years). Participants were randomly assigned to either a *Baseline* condition or an *Intervention* condition. Participants in the *Baseline* condition read that another Mechanical Turk worker in a prior survey had estimated the listing price of a house. Participants again viewed one randomly selected house out of a set of five listings from RedFin.com. Next, participants read that after making a first estimate about the house's listing price, the worker was asked to view the information about the house a second time and make a second estimate. Participants learned that the worker was incentivized for accuracy on both estimation rounds (i.e., that the worker read they would earn five cents if their estimate was within 10% of the correct price). Finally, participants were asked to indicate whether this worker's first or second estimate was more accurate by selecting the appropriate button.

Participants in the *Intervention* condition began the study by completing the same estimation task themselves for a different house. These participants were next asked to indicate whether they thought that their *own* first or second estimate was more accurate. Participants indicated their judgment by selecting the appropriate button. Finally, these participants completed the previously-described task in which they assessed the relative accuracy of a fellow worker's estimates for a different house.

Results

We employed a chi square analysis to examine participants' perceptions of the relative accuracy of their own first and second estimate. This analysis revealed that—unlike individuals' beliefs about others' accuracy documented in our earlier studies—the majority of participants in the *Intervention* condition (60.6%) believed that their *first* estimate was more accurate than their second estimate, $\chi^2 = 13.76$, $p < .001$ (Cohen's d = .43, 95% CI: .20 to .66).

This intervention experience eliminated the previously-documented misprediction of others' accuracy improvement. As in prior studies, the majority of participants in the *Baseline* condition (65.1%) believed that others' second estimates would be more accurate than others' first estimates, a proportion significantly greater than 50% (χ^2 =

27.18, $p < .001$; Cohen's $d = .63$, 95% CI: .39 to .87). However, in the *Intervention* condition only 53.7% of participants believed that others' second estimates were more accurate than others' first estimates, a proportion not significantly different from 50% ($\chi^2 = 1.72$, $p = .19$) and significantly different from the *Baseline* condition, $\chi^2 = 7.62$, $p = .006$ (Cohen's $d = .23$, 95% CI: .07 to .39).

STUDY 6: INTERPERSONAL CONSEQUENCES

Studies 1–5 provide converging evidence that both lay people and experts believe that others' second estimates are more accurate than their first estimates, but that this belief is misguided. Furthermore, after having the experience of making two estimates themselves, individuals seem to conclude that their first estimates, not their second estimates, are actually more accurate. Ironically, this is also not true.

These findings may have important implications for managerial decision-making. Asking others to make second estimates does not simply fail to improve accuracy but may also produce damaging interpersonal consequences. To the extent that estimators themselves believe that their first estimates are more accurate than their second estimates, they may not look kindly upon having to exert additional, likely useless effort. In addition to resenting expending the additional effort that a second estimate requires, individuals may feel that the request signals distrust in their initial estimate (and thus insults their competence). Our final study is designed to document the existence of such interpersonal costs.

To this end, we asked participants to work on an estimation task with a "supervisor" who did or did not request that they make a second estimate. We then captured participants' perceptions of the supervisor. To provide granular insight into the magnitude of the potential interpersonal cost of requesting a second estimate, we compare its effect on supervisor perceptions to the effect of the supervisor allocating progressively less fair payments to the estimator.

Method

We aimed to collect data from 1,750 participants from Mechanical Turk. Participants who failed an attention check at the beginning of the survey were barred from entering the survey, and we collected data from 1,750 participants. Per our pre-registration plan, we removed data from 22 participants who indicated suspicion about the survey's cover story (regarding working with a supervisor), resulting in a final sample of 1,728 participants (mean age = 37; 59% female).

All participants learned that in this study they would work on an estimation task with another Mechanical Turk participant, and that one of them would play the role of "estimator" and the other, the role of "supervisor." Participants further learned that the supervisor would decide which type of estimate the estimator would make (by selecting from the items in Study 1A), and how the accuracy bonus would be divided. They then answered a short set of introductory questions (which they learned would be used to introduce them to their partner), including questions about their first name, age and state of residence.

On the next screen, all participants learned that they had been assigned to the role of estimator. They also learned basic information about their supervisor (i.e., an individual named Mark who was 34 years old and lived in Massachusetts), and that Mark had decided that they would estimate the number of calories in a pictured meal.

Importantly, participants further learned that a 20-cent bonus would be awarded (i.e., a 100% increase in the total payment for participants' survey completion) if their estimate was within 10% of the correct answer. Participants read that the supervisor was in charge of deciding how this bonus would be allocated, and we randomly assigned participants to learn that the supervisor had decided to keep 75%, 70%, 65%, 60%, 55%, or 50% of the bonus for themselves.

Participants then estimated the number of calories in a pictured meal and submitted their estimate to the survey system, which (ostensibly) sent it to their supervisor. They then received a written response. Between conditions, participants saw that their supervisor had either thanked them for their estimate (*Single Estimate* condition), or that their supervisor had thanked them for their estimate and also asked them to make a second estimate of the same quantity (*Second Estimate* condition). To maintain our cover story, we then asked participants to respond to their supervisor by writing a message into an empty field.

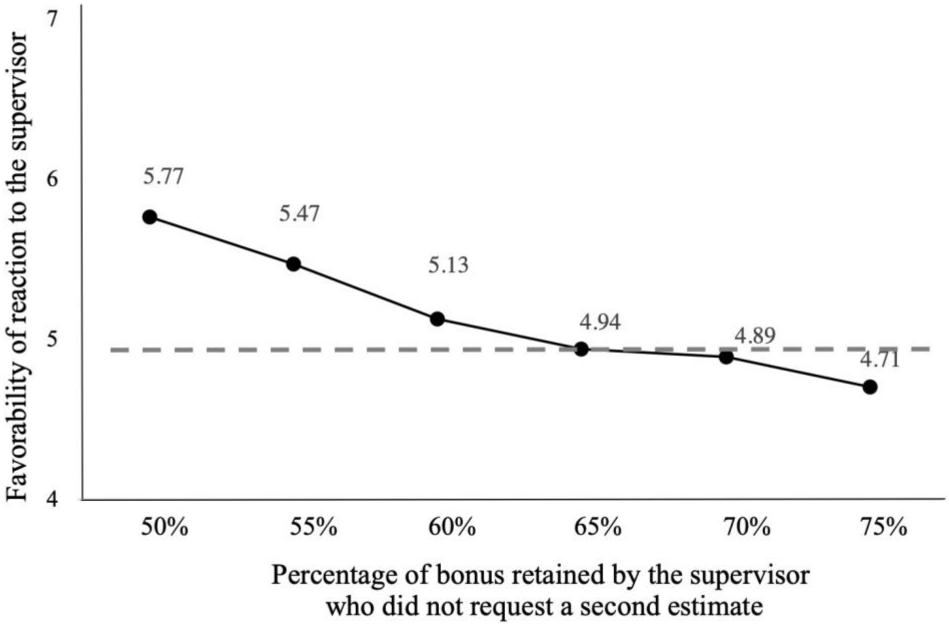
Participants were randomly assigned to one of seven between-participant conditions, which varied in whether or not the supervisor asked for a second estimate and the percentage of the bonus that the supervisor retained for themselves: (1) The supervisor kept 50% of the bonus and asked for a second estimate; (2) The supervisor kept 50% of the bonus and did not ask for a second estimate; (3-7) The supervisor kept increasing shares of the bonus (55%–75%) and did not ask for a second estimate.

Participants then completed a three-item index capturing their evaluation of their supervisor. In particular, participants indicated whether they liked the supervisor, whether they would want to be paired with this supervisor in the future, and whether they would recommend this supervisor to a friend. They indicated their responses to the first question on a 7-point scale (*1: Not at all; 7: Very much*), and their responses to the second two questions on two separate 7-point scales (*1: Definitely no; 7: Definitely yes*). In accordance with our pre-registered analysis plan, we averaged these items into an index capturing participant reactions to their supervisor (Cronbach's $\alpha = .94$).

Results

In the two conditions when the estimator received 50% of the bonus, the supervisor paid an interpersonal cost of requesting a second estimate. Participants evaluated the supervisor less favorably when he asked for a second estimate ($M = 4.94$, $SD = 1.25$) than when he did not ($M = 5.77$, $SD = 1.12$), $t(521) = 8.01$, $p < .001$ (Cohen's $d = .70$, 95% CI: .52 to .88). Importantly, participants preferred the supervisor who kept 55% of the bonus and did not ask for a second estimate ($M = 5.47$, $SD = 1.20$) to one who only kept 50% of the bonus but did ask for a second estimate, $t(482) = 4.78$, $p < .001$ (Cohen's $d = .436$, 95% CI: .26 to .62). Participants still showed a marginally-significant preference for the supervisor who did not ask for a second estimate but kept 60% of the bonus ($M = 5.13$, $SD = 1.36$; $t(508) = 1.70$, $p = .09$; Cohen's $d = .15$, 95% CI: –.02 to .32) relative to the supervisor who did ask for a second estimate. The supervisor who asked for a second estimate and split the bonus equally was rated equivalently to the supervisors who kept 65% of the bonus ($M = 4.94$, $SD = 1.52$; $t(516) = .01$, $p = .99$), and 70% of the bonus ($M = 4.89$, $SD = 1.60$; $t(505) = .37$, $p = .71$), and marginally more favorably than the supervisor who kept 75% of the bonus ($M = 4.71$, $SD = 1.63$; $t(509) = 1.78$, $p = .08$; Figure 4). In other words, the interpersonal costs of requesting a second estimate were equivalent to roughly 20% of the potential bonus payment.

Figure 4. Favorability of reactions to the supervisor as a function of condition.



Notes. The grey dotted line represents reactions to the supervisor who asked for a second estimate and retained 50% of the bonus.

Discussion

Study 6 suggests that the interpersonal cost of requesting a second estimate is not trivial: When compared to the high interpersonal cost incurred by even a small degree of unfair payment (De Cremer, van Dijk, Pillutla 2010; McCall, Steinbeis, Ricard, and Singer 2014; Paz et al. 2017), requesting a second estimate was as damaging as seizing approximately 20% of another participant's (fairly earned) bonus. As noted, the mechanism underlying this interpersonal cost is likely multiply-determined: For example, individuals asked to make second estimates may react negatively to the requester because they recognize that their second estimate will be no better than their first, because they are averse to expending additional effort, and/or because they perceive that the request signals distrust in their initial estimate. Examining these possibilities is a fruitful path for future research. Most relevant to the current work, however, our results suggest that requesting second estimates does not simply fail to improve accuracy, but also may have damaging interpersonal consequences.

GENERAL DISCUSSION

People commonly advise others to reconsider their initial judgments before finalizing them in order to achieve greater accuracy (Cooper 2018; Durnovo 1988; Horner 2020; Shaw 2019; Spiech 2005). The current research investigates the wisdom of this advice. We find that revisiting an estimate does not seem to increase its accuracy, and this failure to improve persists even when people are highly incentivized to do so. Nevertheless, both lay people as well as experts (i.e., judgment and decision-making researchers) seem to expect that second estimates will be more accurate than first estimates. This misprediction persists when observers too are incentivized for correctly gauging relative estimation accuracy.

Importantly, we document people's rationale for their misguided beliefs. Our participants detailed a rich theory about why and how revision improves estimation accuracy. Providing observers information about the specific cognitive processes estimators do (and do not) engage in while reconsidering their estimates attenuates their misguided expectations.

Interestingly, we find that after making two estimates, people believe that their own second estimate is actually inferior to the first. This belief in turn can be leveraged to debias the exaggerated expectations of others' accuracy improvement. Finally, we document interpersonal costs associated with requesting a second estimate.

The current research is the first to systematically examine the potential accuracy benefits of generating a second estimate across a variety of different estimation contexts, incentive structures, elicitation instructions, and estimator populations. In uncovering the gap between estimator performance and observer expectations, we provide the first insight into observers' beliefs about others' second estimates, and the cognitive processes underlying them.

Contributions and Implications

The literature on the "effort heuristic" has documented the perceived benefits of investing effort into a variety of products and services (e.g., Inzlicht, Shenhav, and Olivola 2018; Kruger, Wirtz, Van Boven, and Altermatt 2004). However, in those domains, most individuals can readily distinguish better products and services from worse ones, and effort generally *does* lead to improvement. In the cases where researchers have observed participants erroneously attributing higher quality to equivalent products born from greater effort under experimental conditions, participants had over-applied a generally sensible rule of thumb (e.g., Inzlicht et al. 2018; Kruger et al. 2004).

By contrast, when individuals consider two judgments under uncertainty, it is impossible to accurately judge relative accuracy without having some sense of the correct answer. Furthermore, since our studies show that reconsideration does not improve estimation accuracy, beliefs to the contrary cannot emanate from an overgeneralization of an accurate rule of thumb. Instead, we find these beliefs stem from an imagined set of specific cognitive processes (that either do not happen or do not have the imagined effect). This suggests that the mis-prediction we document originates from a different set of psychological processes than the previously documented effort heuristic.

Interestingly, the literature on the effort heuristic finds that people believe exerting their own effort improves their own outcomes just as it does those of others (Cheng, Mukhopadhyay, and Schrift 2017; Garcia-Rada, John, O'Brien, and Norton 2019; Labroo and Kim 2009; Wan, Rucker, Tormala, and Clarkson 2010). By contrast, we find that when it comes to judgments under uncertainty, people do not perceive that their exertion of additional effort leads to better outcomes, but rather that these judgments get worse upon revision. Thus, the current research uncovers a previously-undocumented self-other asymmetry. Prior work suggests that people's judgments regarding risks (Weinstein 1980, 1984), traits (Alicke et al. 1995), causal attributions (Jones and Nisbett 1972; Malle, Knobe, and Nelson 2007), and abilities (Kruger 1999) can differ as a function of whether the target of judgment is the self or another individual. The current research contributes to this literature by revealing that differences in how people evaluate themselves versus others additionally also shade the perceived utility of revising judgments.

The current phenomenon has potentially important managerial consequences. If people incorrectly believe that others can increase the accuracy of their judgments by simply thinking about them again, then managers may invest employee time into a task that is essentially useless. This investment in fruitless reconsideration may not only leave employees with less time for productive endeavors, but may also carry negative interpersonal consequences. These consequences can be pernicious—indeed, heightened employee dissatisfaction can reduce organizational commitment, increase tardiness, prompt more absenteeism, and increase rates of employee turnover (Kulas, McInnerney, DeMuth, and Jadwinski 2007; Lee 1988; Murphy 1993). The expectation that simply thinking again will improve accuracy may also cause managers to neglect providing employees with the resources and tools that would *actually* make them more accurate. Because good managerial decisions often depend on good estimates and forecasts, this neglect may have important repercussions.

Future Directions

The current research examines whether revisiting an initial judgment improves its accuracy and contrasts these results with the expectations of observers. In addition, we document a moderator of observers' misguided expectations—specifically, that they attenuate when people consider the impact of making a second estimate on their own accuracy. However, there is ample room for further exploration of both mechanisms and boundary conditions.

Necessary conditions for accuracy improvement. Importantly, our findings do not suggest that thinking again about a judgment will *never* improve accuracy. Indeed, it seems highly likely that revisiting a judgment would improve accuracy if additional thought is coupled with appropriate external resources (e.g., feedback and algorithms; Arnold et al. 2016; Kopelman 1986; Remus et al. 1996). Because the current research indicates that it is futile to attempt to achieve greater accuracy by simply reconsidering one's initial judgment, managers should give greater consideration to the resources that do in fact enable accuracy improvement (e.g., feedback, algorithms, and additional information; Arnold et al. 2016; Levine et al. 1975; Kopelman 1986; Remus et al. 1996).

Impediments to accuracy gains in second estimates. We also encourage future research to provide greater insight into why exactly reconsidering initial judgments does not improve accuracy and whether these barriers might be overcome. Studies 3-4 indicated that observers anticipate improvements in estimation accuracy to arise from a set of specific cognitive processes. One possibility is that the belief that others engage in these processes is simply inaccurate. Alternatively, perhaps others do strive to implement these processes, but are not as successful as observers expect. For example, attempts to notice additional discrepant details in the available information may be obstructed by confirmation bias (Bruner et al. 1956; Koriat 1980). Confirmation bias may further thwart people's attempts to recall additional information from memory (Bruner et al. 1956; Koriat et al. 1980; Nickerson 1998). Such attempts may further be rendered ineffective if trying to recall additional information unearths not only true but also false cues (Henkel 2004, 2007; Mcdermott 2006).

Relatedly, attempts to increase accuracy by reweighting cues may be obstructed by anchoring (which may prevent meaningful adjustments from initial estimates; Tversky and Kahneman 1974), the incorporation of additional false recalled details into the weighting scheme, and people's notorious tendency to overweight nonpredictive details (Hall et al. 2007; Kelly and Simmons 2016; Nisbett et al. 1981). As a result, when people are able to escape the clutches of anchoring, they may simply adjust away from one set of nonpredictive details to another.

In sum, if observers are correct in their belief that others making second estimates strive to recall additional relevant information, reweight judgment cues, and notice new details, a variety of impediments may prevent these processes from actually improving estimation accuracy. We encourage future research to examine these possibilities, as well as the manner in which impediments to accuracy improvement may be dismantled.

Conclusion

Accurate judgments are necessary to make good managerial decisions. In attempts to improve judgments' accuracy, people regularly advise others to reconsider their initial judgments before finalizing them (Durnovo 1988; Horner 2020; Shaw 2019; Spiech 2005). The current research investigates the wisdom of this advice. We find that in the domain of making quantitative judgments, trying again yields surprisingly little benefit.

REFERENCES

Alicke M, Klotz M, Breitenbecher D, Yurak T, Vredenburg D (1995) Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology*, 68: 804–825.

ALQahtani D, Rotgans J, Mamede S, ALAlwan I, Magzoub M, Altayeb F, Mohamedani M, Schmidt H (2016) Does time pressure have a negative effect on diagnostic accuracy? *Academic Medicine*. 91(5): 710–716.

Ambady N (2010) The perils of pondering: Intuition and thin slice judgments. *Psychological Inquiry*, 21(4): 271–278.

Ariely D, Au WT, Bender RH, Budescu DV, Dietz CB, Gu H, Wallsten TS, Zauberman G (2000) The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*. 6(2): 130–147.

Arnold S, Stenzel M, Drolet D, Ramachandran G (2016) Using checklists and algorithms to improve qualitative exposure judgment accuracy. *Journal of Occupational and Environmental Hygiene*, 13(3): 159–168.

Barneron M, Allalouf A, Yaniv I (2019) Rate it again: Using the wisdom of many to improve performance evaluations. *Journal of Behavioral Decision Making*, 32(4): 485–492.

Bruner J, Goodnow J, Austin G (1956) A study of thinking. New York: Wiley.

Cheng Y, Mukhopadhyay A, Schrift, R. (2017) Do costly options lead to better outcomes? How the protestant work ethic influences the cost–benefit heuristic in goal pursuit. *Journal of Marketing Research*, 54(4): 636–649.

Cooper J (2018) Developing executives – What's the coaching business case? *Ascendent* <https://the-advisor.co/coaching-business-case/>.

Creyer E, Bettman J, Payne J (1990) The impact of accuracy and effort feedback and goals on adaptive decision behavior. *Journal of Behavioral Decision Making*, 3(1): 1–16.

De Cremer D, van Dijk E, Pillutla M (2010) Explaining unfair offers in ultimatum games and their effects on trust: An experimental approach. *Business Ethics Quarterly*, 107–126.

Durnovo M (1988) A Resource Manual for the Woman in Transition <https://files.eric.ed.gov/fulltext/ED301235.pdf>.

Epley N, Gilovich T (2006) The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, 17(4): 311–318.

Faro D, Rottenstreich Y (2006) Affect, empathy, and regressive mispredictions of others' preferences under risk. *Management Science*, 52(4): 529–541.

Fraundorf S, Benjamin A (2014) Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language*, 71(1): 17–38.

Frederick SW, Mochon D (2012) A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, 141(1): 124–133.

Fujisaki I, Honda H, Ueda K (2018) Diversity of inference strategies can enhance the 'wisdom-of-crowds' effect. *Palgrave Communications*, 4(1): 1–9.

Gaertig C, Simmons J (2020) The Psychology of Second Guesses: Implications for the Wisdom of the Inner Crowd. *Management Science*, Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3668281.

Galinsky A, Mussweiler T (2001) First offers as anchors: the role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, 81(4): 657–669.

Garcia-Rada X, John L, O'Brien E, and Norton M (2019) The revision bias. *Harvard Business School NOM Unit Working Paper*: 19–087.

Halberstad J, Levine G (1999) Effects of Reasons Analysis on the Accuracy of Predicting Basketball Games 1. *Journal of Applied Social Psychology*, 29(3): 517–530.

Hall C, Ariss L, Todorov A (2007) The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes*, 103: 277–290.

Hamlett K (2020) How Are Market Demand, Market Potential & Sales Forecasting Related to Each Other? *Chron* <https://smallbusiness.chron.com/market-demand-market-potential-sales-forecasting-related-other-1485.html>.

Haran U, Moore DA, Morewedge CK (2010) A simple remedy for overprecision in judgment. *Judgment and Decision Making*, 5(7): 467–476.

Harvey N, Harries C, Fischer I (2000) Using advice and assessing its quality. *Organizational Behavior and Human Decision Processes* 81(2): 252–273.

Henkel L (2004) Erroneous memories arising from repeated attempts to remember. *Journal of Memory and Language*, 50(1): 26–46.

Henkel L (2007) The benefits and costs of repeated memory tests for young and older adults. *Psychology and Aging*, 22(3): 580–595.

Herzog S, Hertwig R (2009) The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20(2): 231–237.

Herzog S, Hertwig R (2014) Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, 18(10): 504–506.

Herzog S, Hertwig R (2014) Think twice and then: Combining or choosing in dialectic bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1): 218–232.

Hoffmann J, von Helversen B, Rieskamp J (2013) Deliberation's blindsight: How cognitive load can improve judgments. *Psychological Science*, 24(6): 869–879.

Horner T (2020) IPV Youth League <http://www.iplayvolleyball.net/ipv-youth-leagues.php>.

Horstmann N, Ahlgrimm A, Glöckner A (2009) How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes. *Judgment and Decision Making*, 4: 335–354.

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

Inzlicht M, Shenhav A, Olivola C (2018) The effort paradox: Effort is both costly and valued. *Trends in Cognitive Sciences*, 22(4): 337–349.

Jones E, Nisbett R (1972) The actor and the observer: Divergent perceptions of the cause of behavior. In E. E.Jones, D. E.Kanouse, H. H.Kelley, R. E.Nisbett, S.Valins, & B.Weiner (Eds.), *Attribution: Perceiving the cause of behavior* (pp. 79–94). Morristown, NJ: General Learning Press.

Kahneman D, Tversky A (1972) Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3: 430–454.

Kelly T, Simmons J (2016) When Does Making Detailed Predictions Make Predictions Worse? *Journal of Experimental Psychology: General*, 145(10): 1298–1311.

Kim K (2016) Effect of Dialectical Bootstrapping on Quantitative Estimation. *The International Journal of Creativity and Problem Solving*, 26(1): 43-50.

Kopelman R (1986) Objective feedback. In E. A. Locke (Ed.), Generalizing from laboratory to field settings. Lexington, MA: Lexington Books.

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

Krueger J, Chen L (2014) The first cut is the deepest: Effects of social projection and dialectical bootstrapping on judgmental accuracy. *Social Cognition*, 32(4): 315–336.

Kruger J (1999) Lake Wobegon be gone! The “below-average effect” and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, 77: 221–232.

Kruger J, Wirtz D, Van Boven L, Altermatt T (2004) The effort heuristic. *Journal of Experimental Social Psychology*, 40(1): 91–98.

Kulas J, McInnerney J, DeMuth R, Jadwinski V (2007) Employee satisfaction and theft: Testing climate perceptions as a mediator. *The Journal of Psychology*, 141(4): 389–402.

Labroo A, Kim S (2009) The “instrumentality” heuristic: Why metacognitive difficulty is desirable during goal pursuit. *Psychological Science*, 20(1): 127–134.

Ladesma G (2004) Waste not, want not. *Logistics Today* <https://www.mhlnews.com/global-supply-chain/article/22036633/waste-not-want-not>.

Lee T (1988) How job dissatisfaction leads to employee turnover. *Journal of Business and Psychology*, 2(3): 263–271.

Lehner PE, Adelman L, Cheikes BA, Brown MJ (2008) Confirmation bias in complex analyses. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(3): 584–592.

Levin I, Huneke M, Jasper J (2000) Information processing at successive stages of decision making: Need for cognition and inclusion–exclusion effects. *Organizational Behavior and Human Decision Processes*, 82(2): 171–193.

Levine J, Samet M, Brahlek R (1975) Information seeking with limitations on available information and resources. *Human Factors*, 17(5): 502–513.

Li Z, Rohde K, Wakker P (2017) Improving one’s choices by putting oneself in others’ shoes–An experimental analysis. *Journal of Risk and Uncertainty*, 54(1): 1–13.

Liberman V, Minson JA, Bryan CJ, Ross L (2012) Naïve realism and capturing the “wisdom of dyads” *Journal of Experimental Social Psychology*, 48(2): 507–512.

Logie R, Gilhooly K, Wynn V (1994) Counting on working memory in arithmetic problem solving. *Memory & Cognition*, 22(4): 395–410.

Maki R, Foley J, Kajer W, Thompson R, Willert M (1990) Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4): 609–616.

Malle B, Knobe J, Nelson S (2007) Actor–observer asymmetries in explanations of behavior: New answers to old questions. *Journal of Personality and Social Psychology*, 93: 491–514.

Mata A, Ferreira M, & Sherman S (2013) The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology*, 105(3): 353–373.

McCall C, Steinbeis N, Ricard M, & Singer T (2014) Compassion meditators show less anger, less punishment, and more compensation of victims in response to fairness violations. *Frontiers in Behavioral Neuroscience*, 8: 424.

Mcdermott K (2006) Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, 34(2): 261–267.

Meikle NL, Tenney ER, Moore DA (2016) Overconfidence at work: Does overconfidence survive the checks and balances of organizational life? *Research in Organizational Behavior*, 36: 121–134.

Minson JA, Liberman V, Ross L (2011) Two to tango: Effects of collaboration and disagreement on dyadic judgment. *Personality and Social Psychology Bulletin*, 37(10): 1325–1338.

Moore DA, Healy PJ (2008) The trouble with overconfidence. *Psychological Review*, 115(2): 502–517.

Moxley J, Ericsson K, Charness N, Krampe R (2012) The role of intuition and deliberative thinking in experts' superior tactical decision-making. *Cognition*, 124(1): 72–78.

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

Murphy K (1993) Honesty in the workplace. Pacific Grove, CA: Brooks Cole.

Nickerson R (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2): 175–220.

Nisbett R, Zukier H, Lemley R (1981) The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, 13(2): 248–277.

Obrecht N, Chesney D (2018) Tasks That Prime Deliberative Processes Boost Base Rate Use. *The Annual Meeting of the Cognitive Science Society*.

Oswald ME, Grosjean S (2004) Confirmation bias. *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*, (pp. 79–96). Hove, UK: Psychology Press.

Payne J (1982) Contingent decision behavior. *Psychological Bulletin*, 92(2): 382–402.

Paz V, Nicolaisen-Sobesky E, Collado E, Horta S, Rey C, Rivero M, Berriolo P, Díaz M, Otón M, Pérez A, Fernández-Theoduloz G (2017) Effect of self-esteem on social interactions during the Ultimatum Game. *Psychiatry Research*, 252: 247–255.

Pennycook G, Rand D (2019) Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188: 39–50.

Peterson D, Pitz G (1986) Effects of amount of information on predictions of uncertain quantities. *Acta Psychologica*, 61(3): 229–241.

Remus W, O'Conner M, Griggs K (1996) Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, 66(1): 22–30.

Sah S, Malaviya P, Thompson D (2018) Conflict of Interest Disclosure as an Expertise Cue: Automatic versus Deliberative Processing. In *Academy of Management Proceedings*, 1: 12505.

See, K, Morrison E, Rothman N, Soll J (2011) The detrimental effects of power on confidence, advice taking, and accuracy. *Organizational Behavior and Human Decision Processes*, 116(2): 272–285.

Shaw A (2019) Business Decision Making Process. <https://www.marketingtutor.net/business-decision-making-process/>.

Simmons JP, LeBoeuf RA, Nelson LD (2010) The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology*, 99(6): 917–932.

Singh (2019) Home Renovation Tips. <https://www.lebenstyle.com/interior-design-ideas-house-renovation/>.

Skov RB, Sherman SJ (1986) Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. Journal of Experimental Social Psychology, 22: 93–121.

Slater P. (2010) Smart inventory solutions: Improving the management of engineering materials and spare parts. Industrial Press Inc.

Snyder M, Campbell BH (1980) Testing hypotheses about other people: The role of the hypothesis. *Personality and Social Psychology Bulletin*, 6: 421–426.

Soll JB, Larrick RP (2009) Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3): 780–805.

Spiech S (2005) How to be a great financial analyst. *Strategic Finance*, 86(10): 40–45.

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

Stevenson, M (2018) How to be an ethical leader in 6 steps. *SmartHR* <https://www.smarthr.co.za/how-to-be-an-ethical-leader-in-6-steps/>.

Taber C (2011) Political cognition and public opinion. In R. Y. Shapiro & L. R. Jacobs (Eds.), The Oxford handbook of American public opinion and the media (pp. 368–383). Oxford: Oxford University Press.

Tenney ER, Meikle, NL., Hunsaker, D, Moore, DA, Anderson, C (2019) Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. *Journal of Personality and Social Psychology*, 116(3): 396–415.

Tetlock PE, Mellers BA, Rohrbaugh N, Chen E (2014) Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science*, 23(4): 290–295.

Tordesillas R, Chaiken S (1999) Thinking too much or too little? The effects of introspection on the decision-making process. *Personality and Social Psychology Bulletin*, 25(5): 625–631.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science*, 185: 1124–1130.

Weinstein, N (1980) Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39: 806–820.

Weinstein, N (1984) Why it won't happen to me: Perceptions of risk factors and susceptibility. *Health Psychology*, 3: 431–457.

White, C, Antonakis J (2013) Quantifying accuracy improvement in sets of pooled judgments: does dialectical bootstrapping work? *Psychological Science*, 24(1): 115–116.

Wilson T, Schooler J (1991) Thinking too much: introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60(2): 181–192.

van Dolder D, van den Assem M (2018) The wisdom of the inner crowd in three large natural experiments. *Nature Human Behavior* 2(1): 21–26.

Villamor, F (2020) 6 Vital Strategies to Manage Construction Project Effectively <https://www.fernandovillamorjr.com/business/6-vital-strategies-to-manage-construction-project-effectively/>.

Vul E, Pashler H (2008) Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7): 645–647.

Wan E, Rucker D, Tormala Z, Clarkson J (2010) The effect of regulatory depletion on attitude certainty. *Journal of Marketing Research*, 47(3): 531–541.

Zak A, Waddell B (2010) Simple Excellence: Organizing and Aligning the Management Team in a Lean Transformation. CRC Press.

Ziv I, Leiser D (2013) The need for central resources in answering questions in different domains: Folk psychology, biology, and economics. *Journal of Cognitive Psychology* 25(7): 816–832.

SUPPLEMENTARY STUDY A

CONCEPTUAL REPLICATION OF STUDY 1A

We instructed participants in Study 1A to enter a second estimate that differed from their first in order to ensure that any lack of improvement was not due to participants simply entering the same estimate for both their first and second estimates. However, is it possible that this instruction itself drove the lack of improvement documented in Study 1A? We suspected that this possibility was unlikely, but we investigated it in this supplementary study.

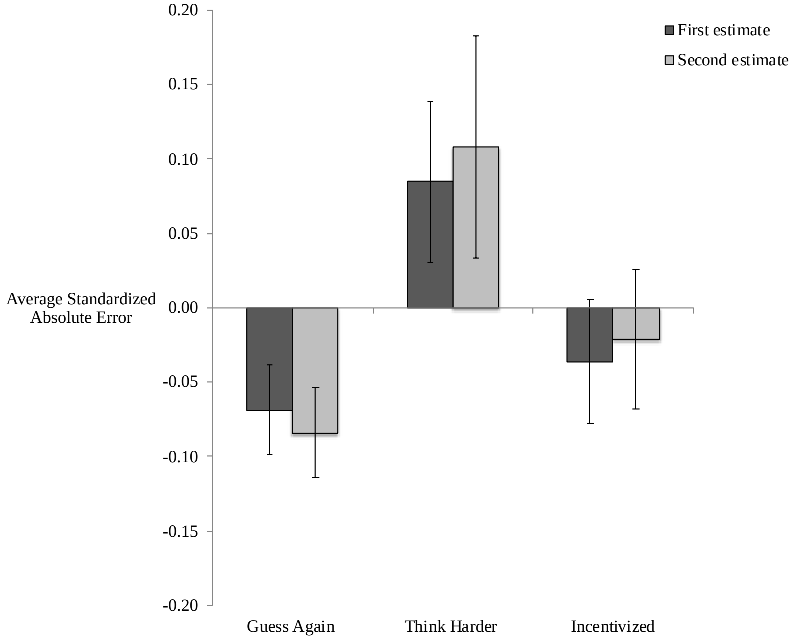
Method

We aimed to recruit 300 participants from Mechanical Turk. Participants who failed an attention check at the beginning of the survey were barred from entering the survey, and we collected data from a total of 300 participants (46.7% female, average age = 38.0 years). The procedures were identical to those in Study 1A, but with one exception—participants were explicitly informed that they could enter their first estimate as their second estimate.

Results

We employed the same methodology detailed in Study 1A to analyze these data. First, as in Study 1A, we followed our pre-registered analysis plan to exclude estimate pairs that differed from each other by an order of magnitude (i.e., by a multiple of ten or more), because we anticipated that such large deviations reflected typographical errors. This exclusion rule resulted in the exclusion of 86 out of the 3,000 estimates (i.e., 2.9% of the estimates). The same regressions employed in Study 1A found no significant decrease in error when participants were incentivized for greater accuracy (b = .02, SE = .02, t = .67, p = .51; 95% CI: –.03 to .06), were prompted to think harder about their estimates (b = .02, SE = .04, t = .61, p = .55; 95% CI: –.05 to .10), or were simply asked to think again (b = –.02, SE = .01, t = –1.90, p = .06; 95% CI: –.03 to .00; Figure S1). Although the marginal effect in the last condition is surprising, we do not detect it in any of our other datasets and thus do not believe it to be reliable. When we collapse the data across conditions, we again see no evidence of accuracy improvement (b = .01, SE = .02, t = .56, p = .58; 95% CI: –.02 to .04).

Figure S1. Descriptive statistics as a function of condition and estimation round in Supplementary Study A.



Note. This figure charts raw means. The error bars are unclustered standard errors.

SUPPLEMENTARY STUDY B

LARGE INCENTIVES ALSO FAIL TO IMPROVE ACCURACY

In this supplementary study, we offered estimators a large bonus for improving the accuracy of their first estimates, in order to test whether the lack of improvement documented in the earlier studies was simply due to lack of motivation. We also employed a new estimation context—estimating others' preferences and behavior as revealed in one of our own surveys.

Method

Stimuli collection: In a pretest, two hundred Mechanical Turk workers responded to five questions presented in Table S1. Participants in Supplementary Study B then estimated the proportion of these pretest participants that gave each response.

Table S1. Estimation targets employed in Supplementary Study B.

Question	Response options	Proportion of participants indicating each response	
Do you currently use a blue tooth brush (i.e., a tooth brush whose primary color is blue)?	Yes vs. No	Yes: 22.5%	No: 77.5%
Is there a plant in the room in which you are currently in?	Yes vs. No	Yes: 35.4%	No: 64.6%
Which of the following foods would you rather have for dinner?	Pizza vs. Chinese food	Pizza: 52.5%	Chinese food: 47.5%
What do you value more in a friendship?	Kindness and generosity vs. complete honesty	Kindness and generosity: 46.5%	Complete honesty: 53.5%
Which do you think would be a more valuable educational experience?	A year at any university you wanted vs. A year leaving in any foreign country you wanted	A year at any university you wanted: 28.6%	A year living in any foreign country you wanted: 71.4%

Notes. Two hundred Mechanical Turk workers responded to the five questions in this table. The proportion of participants that chose each answer is detailed in this table. Participants in Supplementary Study B then estimated the proportion of pre-test participants that gave each response.

Estimation task: Three hundred two participants from Mechanical Turk (58.1% female, average age = 37.4 years) completed an online survey for payment. Participants who failed any one of three attention screeners at the beginning of the survey were barred from entering the survey. All participants viewed the five questions detailed in Table S1. They then estimated the proportion of the prior participants that gave each response. Participants entered their responses into an empty field. To ensure that participants paid at least minimal attention on the first estimation round, we offered them an additional ten cents for each correct estimate that they generated.

As in the prior studies, we then asked participants to make a second estimate for each of the questions that they had previously viewed. This time, participants learned that they would receive an additional five-dollar bonus for each correct estimate that they generated (i.e., a 4,900% increase over the incentive for accurate initial estimates). Participants then entered their second estimates into an empty field. We also asked exploratory items related to participant perceptions of accuracy and whether they sought additional help to make their estimates (see survey materials). None of these items impacted the study's results.

Results

We employed the same analysis strategy detailed in Study 1A, Study 2A, and Supplementary Study A to analyze these data. Specifically, we examined whether making a second estimate reduces error by regressing the z-scored estimation error on estimation round (first estimates = 0; second estimates = 1), clustering the data at the level of participant. In this study, we did not implement the exclusion criteria used in the earlier studies because the estimates in this study were made on a scale from 0 to 100 and we had little concern about participants introducing large typographical errors.

In line with our prior results, we observed no significant increase in accuracy from the first round (where z-scoring forced the mean to be equal to zero) to the second round (M = –.002) of the estimation task (b = –.00, se = .03, t = –.11, p = .91). In sum, we found that a large incentive did not lead participants to improve the accuracy of their estimates upon revision. In Supplementary Study C, we next examined whether incentivized observers who learned about this study anticipate the lack of accuracy improvement we observed.

SUPPLEMENTARY STUDY C

CONVERGING EVIDENCE FOR OBSERVERS' MISPREDICTION

Next, we tested whether incentivized observers who learned about Supplementary Study B anticipated the lack of accuracy improvement we observed.

Method

Two hundred participants from Mechanical Turk (54.0% female, average age = 37.9 years) completed an online survey for payment. Participants who failed any one of three attention screeners at the beginning of the survey were barred from entering the survey.

All participants read details about Supplemental Study B, including that other Mechanical Turk workers had estimated their fellow workers' responses to five different questions. On the next screen, participants viewed one randomly-selected question from these five questions, and learned about the incentives for accuracy that we offered to the estimators on both rounds of the task.

Mirroring Study 2B's design, participants next learned that one of the participants from Supplementary Study B had been randomly assigned to be their partner, and they had the opportunity to receive a bonus based on the accuracy of their partner's estimates. Participants then learned that they could choose whether to base this bonus on their partner's first estimate or second estimate, and that they would receive a five-cent bonus if their chosen estimate was within 1% of the accurate answer. Participants entered their choice by selecting a button labeled either "My partner's first estimate" or "My partner's second estimate."

Results

As in Study 2B, chi-square analysis revealed that participants were more likely to choose their partner's second estimate (63.5%) than their partner's first estimate (36.5%), $\chi^2(df = 1, N = 200) = 14.58$, $p < .001$ (Cohen's $d = .56$, 95% CI: .27 to .85). Thus, when participants were directly incentivized to choose the estimate that was most accurate, they were more likely to choose others' second estimates than others' first estimates.

SUPPLEMENTAL MATERIALS D

ADDITIONAL DETAILS REGARDING STUDY 3

Participants who indicated that the first estimate was more accurate chose between the following four categories: “When people make their second estimate, they may notice irrelevant new details in the information that they did not notice when they had made their first estimate”; “When people make their second estimate, they may remember irrelevant information that did not come to mind when they had made their first estimate”; “When people make their second estimate, they may second-guess their gut feeling and weigh the information less effectively than they did when they had made their first estimate”; “Another reason not listed above.” Participants who chose the last option were provided with an empty field to describe their reason’s category.

Among participants who believed that first estimates were most accurate, 17.3% categorized their rationale as describing the intuition that second estimates are worsened by irrelevant new information noticed during the second estimate, 10.7% categorized their rationale as describing the intuition that second estimates are worsened by irrelevant new information remembered during the second estimate, 61.3% categorized their rationale as describing the intuition that second estimates are worsened by less effective weighting of information during the second estimate, and 10.7% categorized their rationale into a different category.