Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Known Unknowns: A Critical Determinant of Confidence and Calibration

Daniel J. Walters, Philip M. Fernbach, Craig R. Fox, Steven A. Sloman

To cite this article:
Daniel J. Walters, Philip M. Fernbach, Craig R. Fox, Steven A. Sloman (2017) Known Unknowns: A Critical Determinant of Confidence and Calibration. Management Science 63(12):4298-4307. https://doi.org/10.1287/mnsc.2016.2580

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 12, December 2017, pp. 4298–4307
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Known Unknowns: A Critical Determinant of Confidence and Calibration

Daniel J. Walters,[a] Philip M. Fernbach,[b] Craig R. Fox,[a] Steven A. Sloman[c]

[a] Anderson School of Management, University of California, Los Angeles, Los Angeles, California 90024; [b] Leeds School of Business, University of Colorado Boulder, Boulder, Colorado 80309; [c] Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, Rhode Island 02912

Contact: daniel.walters@insead.edu (DJW); philip.fernbach@gmail.com (PMF); craig.fox@anderson.ucla.edu (CRF); steven_sloman@brown.edu (SAS)

Received: February 20, 2015
Revised: January 26, 2016; June 2, 2016
Accepted: June 7, 2016
Published Online in Articles in Advance: December 16, 2016

https://doi.org/10.1287/mnsc.2016.2580

Abstract. We propose that an important determinant of judged confidence is the evaluation of evidence that is unknown or missing, and overconfidence is often driven by the neglect of unknowns. We contrast this account with prior research suggesting that overconfidence is due to biased processing of known evidence in favor of a focal hypothesis. In Study 1, we asked participants to list their thoughts as they answered two-alternative forced-choice trivia questions and judged the probability that their answers were correct. Participants who thought more about unknowns were less overconfident. In Studies 2 and 3, we asked participants to list unknowns before assessing their confidence. "Considering the unknowns" reduced overconfidence substantially and was more effective than the classic "consider the alternative" debiasing technique. Moreover, considering the unknowns selectively reduced confidence in domains where participants were overconfident but did not affect confidence in domains where participants were well-calibrated or underconfident.

History: Accepted by Yuval Rottenstreich, judgment and decision making.
Funding: This work was supported in part by a grant from the Thrive Center for Human Development and John Templeton Foundation project on the Science of Intellectual Humility.
Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mnsc.2016.2580.

Keywords: behavioral decision making • overconfidence • calibration • judgment • decision analysis

In the run-up to the Iraq war of 2003 many leaders in the United States expressed great confidence that the Iraqi President, Saddam Hussein, was developing weapons of mass destruction (WMDs). In a letter sent to President George W. Bush in 2001, 10 of the most influential congressman, both Democrats and Republicans, wrote "There is no doubt that . . . Saddam Hussein has reinvigorated his weapons program. Reports indicate that biological, chemical and nuclear programs continue apace and may be back to pre-Gulf War status."[1] Senator Jay Rockefeller expressed the same sentiment in a 2002 speech: "There is unmistakable evidence that Saddam Hussein is working aggressively to develop nuclear weapons and will likely have nuclear weapons within the next five years."[2] As we now know, there were no WMDs, so these statements expressing "no doubt" and "unmistakable evidence" apparently reflected overconfidence that had major geopolitical consequences. While this example may be extreme, it is not unusual. Overconfidence has been implicated in a wide range of decision errors, from going to war (Johnson 2004) to treatment of medical conditions (Baumann et al. 1991, Oskamp 1965) to corporate investments (Malmendier and Tate 2005) to market entry (Camerer and Lovallo 1999, Mahajan 1992).

A great deal of research has attempted to understand the sources of error in judging confidence with an eye to developing debiasing techniques. Much of this research has attributed overconfidence to a systematic tendency to seek or overweight known evidence for a favored hypothesis over its alternatives. In the case of the Iraq war, overconfidence may have been driven in part by the Bush administration's promotion of the hypothesis that Iraq was developing WMDs and the bias among observers to seek and overweight evidence confirming this hypothesis. An abundance of research has found that people tend to focus disproportionately on evidence for a focal hypothesis relative to alternatives (Koriat et al. 1980, Hoch 1985, Klayman 1995), and that they tend to seek evidence consistent with the focal hypothesis as part of a positive test strategy (Mynatt et al. 1977, Klayman and Ha 1987, Nickerson 1998), wishful thinking (Babad 1987), motivated reasoning (Kunda 1990), or protection of their self-image from failure and regret (Larrick 1993). One reason this approach to understanding overconfidence has been so influential is because it has led to successful debiasing techniques that tend to improve judgment calibration. Overconfidence can be reduced by prompting

people to "consider the alternative" (Koriat et al. 1980) or by designating a member of a decision-making team to advocate for the alternative ("devil's advocate technique"; Schwenk and Cosier 1980).

A second class of theories of confidence represents the mapping between balance of known evidence and judged probabilities. Griffin and Tversky (1992) distinguish strength of evidence (i.e., balance) from weight of evidence (i.e., reliability or diagnosticity). They argue that when judging probabilities, people tend focus on strength of evidence and give insufficient regard to weight of evidence. This can contribute to both overconfidence (when strength of evidence is high and weight of evidence is low) and underconfidence (when strength of evidence is low and weight of evidence is high). People focus on strength of evidence while neglecting weight of evidence because they overestimate the predictive validity of evidence that is representative (Tversky and Kahneman 1974), internally consistent (Kahneman and Tversky 1973), and based on small samples (Tversky and Kahneman 1971). Similarly, in support theory (Rottenstreich and Tversky 1997, Tversky and Koehler 1994), probability is determined by the perceived balance of evidence for a hypothesis relative to its alternative. Overconfidence can occur due to scaling the perceived balance of evidence to overly extreme judged probabilities (see Fox 1999), for instance when perceived evidence is seen as especially predictive of outcomes (Tannenbaum et al. 2017), or when the environment does not provide particularly diagnostic cues (Brenner et al. 2005). In evidence accumulation models, confidence is determined by weighting evidence based on feeling (Ferrell and McGoey 1980) or self-consistency (Koriat 2012), and overconfidence can occur when these cues are overestimated.

We propose that when assessing confidence, people may also look directly to specific pieces of *unknown evidence* to determine how to weight or scale the balance of known evidence. By unknown evidence, we mean a variable the value of which is unknown but if it were known should change one's level of confidence. For instance, prior to the invasion of Iraq, Saddam Hussein's motivation for not cooperating with weapons inspectors was unknown to most American observers. Mr. Hussein may have wanted the world to believe that he *did possess WMDs* (to increase the perceived strength of the Iraqi military) or that he *did not possess WMDs* (to reduce the likelihood of a U.S.-led invasion). Becoming aware of this important unknown factor would not change the information available to a judge. However, awareness of the unknown is likely to decrease confidence by making the judge aware that he or she is missing critical information. Unknown evidence can potentially support the focal or an alternative hypothesis once the unknown is resolved. So being aware of more unknown evidence should generally lead to less extreme confidence in both outcomes.

Biased evaluation of known evidence clearly plays a role in overconfidence, but failure to adequately consider unknowns may be equally important. A growing body of literature shows that people tend to think the world is simpler and more predictable than it is because they focus on what they know and tend to neglect what they do not know. For instance, people tend to think they understand various types of causal systems, from machines to public policies, in much greater detail than they actually do (Alter et al. 2010, Fernbach et al. 2013, Rozenblit and Keil 2002). People also tend to neglect unknown causes of system failure when diagnosing problems such as why a car won't start (Fischhoff et al. 1978), and they underestimate the possibility of unknown or unexpected delays in the planning fallacy (Buehler et al. 1994). People also exhibit a "censorship bias" in which they fail to account for missing sample information when forming beliefs about an underlying population (Feiler et al. 2013). Similarly, consumers tend to neglect unknown or unmentioned attributes when evaluating products (Sanbonmatsu et al. 1991, 1992). More generally, Kahneman (2011) uses the focus on known relative to unknown information as an organizing principle for many phenomena in judgment and decision making, which he refers to as the "What You See Is All There Is" (WYSIATI) principle.

We have proposed that judged confidence depends in part on the judge's assessment of how much evidence is missing or unknown. In particular, we predict that greater appreciation of unknowns will be associated with judged probabilities that tend more toward the "ignorance prior" probability of $1/n$ in an n-alternative forced-choice paradigm (e.g., $\frac{1}{2}$ when there are two alternatives) whereas less appreciation of unknowns will be associated with more extreme confidence judgments that depart more from the ignorance prior. Consistent with this hypothesis, previous studies suggest that when people are less knowledgeable, they provide less extreme probability judgments. Fox and Clemen (2005) report that judged probabilities of n exclusive and exhaustive events—for example, the branches from a chance node in a decision tree—were biased more strongly toward probabilities of $1/n$ for events about which participants had less knowledge or expertise. Likewise, See et al. (2006) found that judged probabilities were biased more strongly toward $1/n$ when participants had less opportunity to learn the frequencies of observed events or when they reported feeling less confident in what they had learned.

In Study 1, we use a correlational, thought-listing paradigm to test whether differences in consideration of unknowns predict differences in confidence and overconfidence, controlling for the balance of known

evidence. We also examine whether underappreciation of unknowns is associated with overconfidence. In particular, we predict that prompting people to consider unknowns will reduce overconfidence. In Studies 2 and 3, we introduce a novel debiasing technique, "consider the unknowns" (CTU), in which participants are asked to reflect on what they do not know before reporting their confidence, and we compare the efficacy of this technique to the classic "consider the alternative" intervention (Koriat et al. 1980).

Study 1

We asked participants to judge the probability of making a correct choice in a two-alternative forced-choice (2AFC) task involving general knowledge questions. The 2AFC paradigm is a well-studied context in which people often exhibit overconfidence (for reviews, see McClelland and Bolger 1994, Koehler et al. 2002, Griffin and Brenner 2004). As participants completed the task, we also asked them to provide reasons for their judgments using a thought-listing procedure (Johnson et al. 2007). We then asked participants to self-code each of their reasons on the extent to which it referred to known versus unknown evidence. In addition, we asked two hypothesis-blind judges to code the extent to which each reason supported the chosen or alternative option. We predicted that respondents would exhibit lower confidence to the extent that they thought about more unknown evidence and that this relationship would hold after controlling for the balance of known evidence.

Methods

We recruited 134 students at the University of Colorado Boulder to participate in a laboratory experiment in exchange for a $3 payment (49% female; mean age = 20.0). We first asked them to answer 10 2AFC questions, each with two possible answers adapted from Klayman et al. (1999); a complete set of questions is provided in Online Appendix A. After answering each question, we asked participants to report their confidence by estimating the probability that they correctly answered the question, on a scale from 50% to 100%.

For the first 3 of 10 questions (questions 1–3 in Online Appendix A), we asked participants to list the reasons for their confidence:

> As you answer the question, please think of all the reasons that make you {more/less} confident you know the answer and all the reasons that make you {less/more} confident. We will ask you to enter your reasons one at a time. Type your first complete reason in the box below and, as soon as you are done, hit the "enter" key to submit it. You may enter your reasons in any order.

The order of the words "more" and "less" was randomly determined for each participant and had no effect on confidence or answer choice. Participants could list as many or as few reasons as came to mind. The entered reasons then appeared, and participants had an opportunity to enter more reasons. Participants listed reasons while viewing the 2AFC question, and they could change both their answer and confidence while listing reasons.

After completing all 10 questions, we reminded participants of each of the reasons they provided for the first three questions. We then asked them to rate each reason as being about known or unknown evidence on a seven-point scale (1 = completely known; 7 = completely unknown). We explicitly asked participants to rate how known versus unknown the reason was rather than how much each reason improved the participant's estimate in order to make sure we were measuring the content of the reason, rather than the effect of the reason on confidence. A sample of the rating instructions can be found in Online Appendix B. Finally, we collected demographic data and debriefed participants.

Results

Unknown Rating and Reasons Generated. For the three question for which participants provided and rated reasons for their confidence estimates, participants provided an average of 2.36 reasons per question with an interquartile range of (2.35, 2.56). We calculated participants' average rating of reasons for how much they involved unknown evidence (1 = completely known; 7 = completely unknown). The mean rating was 3.45 with an interquartile range of (2.56, 5.33), and 63% of participants had an average rating below the scale midpoint, suggesting that most participants reported more known than unknown evidence. Reasons rated as known tended to be statements of facts, whereas reasons rated as unknown tended to be statements about missing information or lack of relevant knowledge. Online Appendix C provides examples of representative known and unknown reasons generated by participants.

Confidence, Percent Correct, and Overconfidence. Across the three questions where reasons were provided, mean confidence ratings were 67.4%, while on average participants answered 62.2% of questions correctly. For each participant, we calculated overconfidence following conventional methods (see McClelland and Bolger 1994, Koehler et al. 2002, Griffin and Brenner 2004) by subtracting the percentage of all items answered correctly from average confidence, resulting in mean overconfidence of 5.2%, significantly more than 0%, $t(133) = 2.36$, $p < 0.05$, replicating previous work (e.g., Koriat et al. 1980). Confidence, percent correct, and overconfidence did not vary significantly for the seven questions where no reasons were provided compared to the three where reasons were provided.

We next examined the relationship between unknown ratings, confidence, percent correct, and overconfidence across the three questions for which participants provided reasons. We first calculated the average confidence and percent correct on these question. We regressed the average confidence judgment on the average unknown rating. As we predicted, participants who provided reasons that they rated as more unknown were less confident, $b = -3.11$, 95% confidence interval (CI) $[-4.63, -1.59]$, $p < 0.001$. We also regressed percent correct on the known versus unknown rating and found no significant relationship, $b = 0.66$, 95% CI $[-2.79, 4.11]$, $p > 0.5$. We then regressed overconfidence on unknown ratings. Participants who generated reasons that they rated as more unknown exhibited less overconfidence, $b = -3.77$, 95% CI $[-7.38, -0.16]$, $p < 0.05$. To assess the level of unknown rating at which overconfidence becomes significant, we conducted a floodlight analysis (Spiller et al. 2013). The Johnson–Neyman point occurred at an unknown rating of 3.1, meaning that at this level of average unknown rating and above it, overconfidence did not significantly differ from 0. Below this average unknown rating, participants were significantly overconfident. At no level of average unknown rating were participants underconfident.

Balance of Known Evidence. We asked two hypothesis-blind coders to code participants' reasons according to the extent to which they appear to support the chosen versus alternative option, using a seven-point scale (1 = strong support of alternative option; 7 = strong support of the chosen option). Coders were not provided with the unknown rating or any other data besides the study questions and participant reasons. Nine participants did not provide reasons on at least one of the questions and were not scored by coders. Inter-rater reliability of these scores was high (Cronbach's $\alpha = 0.80$). Not surprisingly, mean balance of known evidence was 5.33 in favor of the chosen option, with an interquartile range of (4.81, 5.58). Online Appendix C provides examples of representative reasons coded as supporting the chosen and the alternative options. Rated support was not significantly correlated with unknown rating ($r = -0.12$, $p = 0.201$). Focusing only on the questions where participants provided and self-coded reasons, we ran three separate regressions with balance of known evidence as the independent variable and confidence, percent correct, or overconfidence as the dependent variable. Participants who provided reasons that were rated as more supportive of the focal compared to alternative hypothesis were marginally more confident in their choices $b = 2.85$, 95% CI $[-0.53, 6.24]$, $p = 0.098$. Balance of known evidence did not significantly predict percent correct, $p > 0.1$, or overconfidence, $p > 0.5$.

We next conducted hierarchical regressions with average confidence across the three questions for which participants provided reasons as the dependent variable, and known versus unknown rating and balance of known evidence as the predictors. The model R-squared increased from 0.02 to 0.15 when adding known versus unknown rating to balance of known evidence, $F(1, 122) = 18.15$, $p < 0.0001$. When adding balance of known evidence to known versus unknown rating, the R-squared marginally increased, from 0.11 to 0.15, $F(1, 122) = 3.68$, $p = 0.057$. This is consistent with our hypothesis that known unknowns contribute to confidence in addition to the balance of known evidence for the chosen versus alternative option.

Within-Participants Analysis. Because each participant rated multiple items, we were also able to perform a within-participant analysis to examine if an individual's confidence, percent correct, and/or overconfidence varied as he or she listed reasons that were more unknown across different questions. For each participant, we examined the relationship between question-level known versus unknown rating and confidence, accuracy, and overconfidence. For each of the three questions, we recorded judged confidence and unknown rating. We scored accuracy as a 1 if correct and a 0 if incorrect, and scored overconfidence as confidence minus accuracy. To analyze the data, we used a linear regression with unknown rating for a particular question as the independent variable and confidence as the dependent variable while clustering standard errors by participant. Replicating the between-participant analysis, participants were less confident when they provided more unknown reasons, $b = -3.73$, 95% CI $[-4.45, -3.00]$, $p < 0.001$. Next, we ran the same regression with overconfidence as the dependent variable. Again replicating the between-participant analysis, higher unknown ratings were related to less overconfidence, $b = -6.97$, 95% CI $[-9.62, -4.32]$, $p < 0.001$. Finally, we ran the same regression with percent correct as the dependent variable. Interestingly, higher unknown ratings significantly predicted percent correct, $b = 3.25$, 95% CI $[0.59, 5.90]$, $p < 0.05$, a result that we did not predict ex ante.

Discussion

This study showed that appreciation of unknowns is related to both confidence and overconfidence. Focusing on more known evidence was associated with greater overconfidence, whereas generating reasons that were rated as entailing more unknown evidence was associated with less overconfidence. Previous research has attributed confidence primarily to the processing of the balance of known evidence. Unknown ratings significantly predicted confidence after controlling for the balance of known evidence, suggesting that consideration of unknowns also contributes to judged confidence.

While the results of Study 1 support our hypothesis concerning the role of known unknowns, we acknowledge that the evidence is correlational and thus open to alternative interpretations. For instance, it is possible that those who felt less confident were more likely to reference unknowns rather than the other way around. In Studies 2 and 3, we experimentally manipulate consideration of unknowns to provide causal evidence of the determinants of overconfidence.

Study 2

In Study 2, we manipulated thinking about unknowns by explicitly asking some participants to "consider the unknowns" (CTU), and we compared the effectiveness of this intervention to the classic "consider the alternative" (CTA) debiasing intervention, in which people are asked to consider known evidence for the alternative hypothesis (Koriat et al. 1980). Considering the alternative has been shown to reduce overconfidence, in part by increasing the percent correct. For example, Koriat et al. (1980) found that percent correct in the control condition was 62.9% compared to 69.7% when people were asked to consider the alternative in the 2AFC paradigm. We believe that as people consider the alternative, they sometimes correctly realize that there is more evidence in favor of the alternative and switch their choice. Thus, considering the alternative can increase percent correct and decrease confidence. In contrast, considering the unknowns should reduce overconfidence only by reducing misplaced confidence and should not cause people to switch their choice.

Methods

We recruited 254 participants at the University of California, Los Angeles, from an online university subject pool to participate in a laboratory experiment in exchange for $3 dollars plus a performance incentive (75.7% female; mean age = 21.0). The performance incentive could range up to $212 (see Online Appendix D for details).

Participants assessed their confidence that they provided the correct answer to each of eight general knowledge questions in a four-alternative forced-choice (4AFC) format. A complete list of questions is displayed in Online Appendix E. We randomly assigned participants to one of three conditions: no treatment, CTA, and CTU. In the no treatment condition, participants answered the questions and estimated their confidence without providing any additional information. In the CTA condition, we adapted the procedure from Koriat et al. (1980) in which participants in a 2AFC paradigm were prompted to list reasons supporting the nonchosen option (the alternative hypothesis) before making a confidence judgment. In our study, we asked participants to generate reasons supporting one of three possible nonchosen options:

> Write down in the spaces provided two reasons that support one of the alternative choices (nonchosen options). Please write the best reasons you can think of that provides evidence for the options you have rejected. For example, in answering the question: 'Which of these cars has a larger engine by volume: Mitsubishi Lancer, Nissan Altima, Mazda CX-5, or Subaru Impreza?' If you chose 'Nissan Altima' you would then list reasons that the correct answer might be the Lancer, the CX-5 or the Impreza.

In the CTU condition we asked participants to:

> Write down in the space provided two pieces of missing information or two unknown factors that would help you determine the correct choice, if known. For example, in answering the question: 'Which of these cars has a larger engine by volume: Mitsubishi Lancer, Nissan Altima, Mazda CX-5, or Subaru Impreza?' An unknown might be: 'I don't know what a CX-5 is,' or 'I don't know if a Lancer is a sedan or an SUV.' What's important is that you write down two factors that are unknown to you.

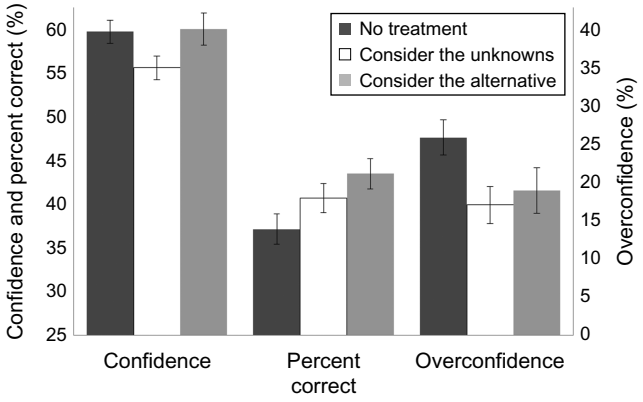
Online Appendix C shows examples of representative reasons generated by participants in the CTU and CTA conditions.

Results

Figure 1 displays the mean level of confidence, percent correct, and overconfidence across the three conditions. Confidence was calculated as the average level of confidence across all eight questions for each participant, percent correct was calculated as the percent correct across all eight questions, and overconfidence was calculated as the difference between the two.

We first analyzed the two treatment conditions against the no treatment condition and against each other. Participants in the CTU condition exhibited

Figure 1. Percent Correct, and Overconfidence in the No Treatment, CTU, and CTA Conditions



Notes. Confidence and percent correct are shown on the left vertical axis and overconfidence is shown on the right vertical axis. Standard errors displayed.

lower confidence than those in the no treatment condition, 56.8% versus 61.1%, $t(170) = 2.14$, $p < 0.05$, and marginally lower confidence than those in the CTA condition, 61.4%, $t(166) = 1.96$, $p = 0.052$. Confidence in the CTA condition did not differ significantly from the no treatment condition, $t(166) < 1$, not significant.

Percent correct in the CTU condition was not significantly different than in the no treatment condition, 41.3% versus 37.6%, $t(170) = 1.46$, $p > 0.1$, or the CTA condition, 44.2%, $t(166) = 1.18$, $p > 0.1$. Percent correct in the CTA condition was significantly higher than the no treatment condition, $t(166) = 2.59$, $p = 0.01$.

Overconfidence in the CTU condition was significantly lower than in the no treatment condition, 15.5% versus 23.5%, $t(170) = 2.60$, $p = 0.01$, and was no different than in the CTA condition, 17.2%, $t(166) < 1$, not significant. Overconfidence in the CTA condition was marginally lower than in the no treatment condition, $t(166) = 1.82$, $p = 0.070$.

Discussion

Considering the unknowns reduced confidence, resulting in decreased overconfidence relative to the no treatment condition. In contrast, considering the alternative did not reduce confidence but did improve percent correct, resulting in marginally less overconfidence than the no treatment condition. Thus, both debiasing techniques showed some efficacy, but considering the unknowns was more effective at reducing confidence.

One limitation of Studies 1 and 2 is they do not distinguish whether considering the unknowns generally improves calibration (i.e., meta-knowledge concerning one's accuracy) or whether it merely reduces confidence on questions where people are already overconfident. The downside of a general reduction in confidence is that where people are ordinarily well-calibrated it would lead to underconfidence, and where people are ordinarily underconfident it would exacerbate this bias. Study 3 allows us to examine the extent to which improvements in calibration following the CTU intervention reflect a nonspecific reduction in confidence versus selective adjustment when confidence is misplaced.

Study 3

We designed Study 3 to replicate and extend the results of Study 2 by enhancing the design in three respects. First, to address the possible concern that the questions in Study 2 may have been especially difficult, which can lead to overconfidence through unbiased judgment error (Erev et al. 1994, Gigerenzer et al. 1991, Soll 1996), we randomly generated the questions from a database of 778,169 questions across nine domains provided to us by Jack Soll (personal communication, November, 2013). Second, in Study 3, we used a within-participant comparison between control and treatment to generalize the results beyond the between-participant design of Study 2. Finally, to establish the generality of the effects, Study 3 relies on a 2AFC paradigm, whereas Study 2 used 4AFC.

Random stimulus sampling and the 2AFC format provide an additional benefit. Because we expect baseline overconfidence to vary across domains (see Klayman et al. 1999), Study 3 allows us to examine the extent to which improvements in calibration due to the CTU prompt are driven by a general reduction in confidence or selective adjustments that depend on the degree of misplaced confidence. If CTU instead has a selective effect, it can provide a more useful and informative method for reducing confidence. To test this, we compare changes in overconfidence in domains where participants are normally overconfident versus those where they are normally well-calibrated or underconfident.

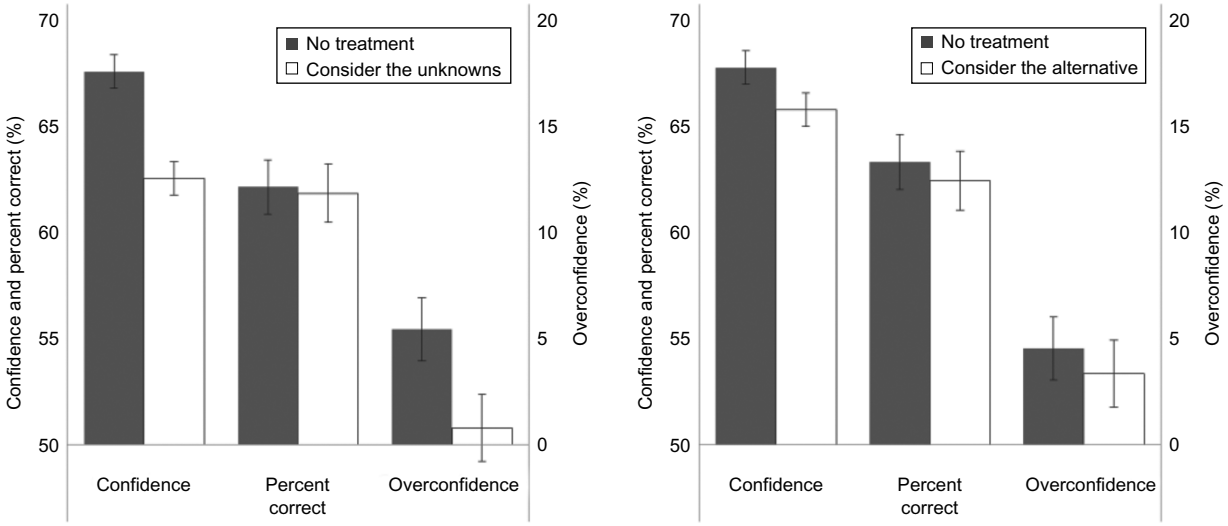
Methods

We recruited 270 participants through a Qualtrics panel in exchange for $4 (66.3% female; mean age = 49.2). One participant did not finish the study and 19 participants (7%) requested that their data not be used in an opt-out option in the study debrief, leaving a sample size of 250.

Participants answered 20 general-knowledge questions in a 2AFC format and assessed their confidence that they provided the correct answer. For each question, we asked participants to pick the correct answer and assess their confidence on a scale from 50% to 100%. The 20 questions were grouped into two blocks of 10 questions each: the first block was the no treatment block and the second block was the treatment block. Before the first block, participants read a brief set of instructions, completed a practice problem, and completed the 10 questions with each question presented on a separate screen. Next, participants were randomly assigned to either the CTA or CTU treatment condition. Depending on condition, participants read instructions similar to CTA or CTU conditions used in Study 2 and completed the second block of questions, this time elaborating on either the alternative or unknowns for each question, following the procedure of Study 2. Online Appendix C shows examples of representative reasons generated by participants in the CTU and CTA conditions.

Each participant received a randomly selected sample of questions drawn from a population of 778,169 question combinations developed by Jack Soll and colleagues. A complete list of question domains is displayed in Online Appendix F. Prior to the study, we created all possible question combinations and then randomly selected five questions per domain, for a total of 45 questions. Each participant received 20 of these questions, sampled at random without replacement, following a method similar to Klayman et al. (1999).

Figure 2. Confidence, Percent Correct, and Overconfidence for Questions With and Without Treatment



Notes. Confidence and percent correct are shown on the left vertical axis, and overconfidence is shown on the right vertical axis. Left panel: CTU. Right panel: CTA. Standard errors are displayed.

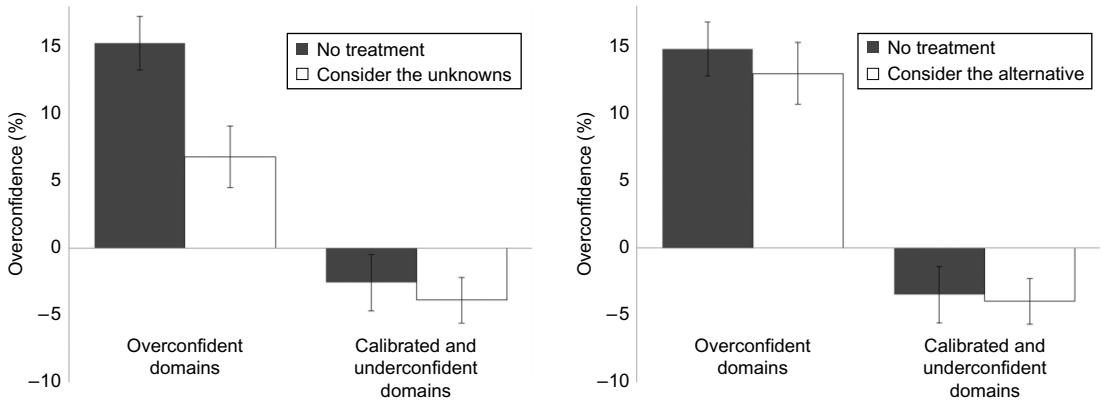
Results

Figure 2 displays mean confidence, percent correct, and overconfidence in the CTU and CTA conditions for the first 10 questions, where there was no treatment, and the last 10 questions, where participants considered the unknowns or the alternative. Replicating Study 2, considering the unknowns reduced confidence and overconfidence and in this case clearly had no effect on percent correct. In line with Study 2, considering the unknowns was more effective at reducing confidence than considering the alternative. For participants in the CTU condition, confidence was lower after generating unknowns than when answering the questions with no treatment, 62.8% versus 67.8%, $t(120) = 6.14$, $p < 0.001$. For participants in the CTA condition, confidence was also slightly lower after generating alternatives than when answering the questions with no treatment, 66.0% versus 68.0%, $t(128) = 2.37$, $p < 0.05$. However, the effect of considering the unknowns on confidence was larger than considering the alternative, $t(248) = 2.55$, $p = 0.01$. Considering the unknowns also reduced overconfidence relative to no treatment, from 5.5% to 0.8%, $t(120) = 2.70$, $p < 0.01$, whereas considering the alternative did not significantly reduce overconfidence, from 4.5% to 3.4%, $t(128) < 1$, $p > 0.5$. While the reduction in overconfidence in the CTU condition (4.7%) was greater than in the CTA condition (1.1%), this difference did not reach statistical significance, $t(248) = 1.38$, $p = 0.16$. However, overconfidence was not statistically different from 0 after considering unknowns, $t(134) < 1$, $p > 0.5$, whereas after considering the alternative, overconfidence persisted, $t(128) = 2.30$, $p < 0.05$. Unlike in Study 2, neither manipulation significantly affected percent correct (means for CTU versus no treatment = 62.0% versus 62.3%, $p > 0.5$; means for CTA versus no treatment = 62.6% versus 63.5%, $p > 0.5$).

We next examined whether considering the unknowns had a larger effect on answers where participants are normally more overconfident. We first identified domains for which participants exhibited statistically significant overconfidence and domains for which they exhibited calibrated or underconfident judgment. We identified domains using a split-sample method similar to Klayman et al. (1999) so that we could rule out regression to the mean as a trivial mechanism driving improvement (see Online Appendix G for additional details). Participants were overconfident in four domains (president elected first, food calories, beverage calories, and atomic weight) and calibrated or underconfident in five domains (country life expectancy, distance from Kansas City, state populations, movie box office revenue, and car miles per gallon). For each participant, we computed four overconfidence scores: (1) overconfident domains with a treatment, (2) overconfident domains without a treatment, (3) calibrated/underconfident domains with a treatment, and (4) calibrated/underconfident domains without a treatment (see Figure 3).

We analyzed the CTU and CTA conditions separately using within-participant regression models, with overconfidence as the dependent variable. The independent variables were domain type (overconfident versus calibrated/underconfident), treatment (treatment versus no treatment), and their interaction. In overconfident domains, overconfidence was lower after considering the unknowns than when answering the questions with no treatment, 6.8% versus 15.3%, $b = 8.5$, 95% CI [3.2, 13.9], $p < 0.01$. In contrast, in calibrated/underconfident domains,

Figure 3. Overconfidence on Questions With and Without Treatment in Overconfident Domains and Calibrated/Underconfident Domains



Notes. Left panel: CTU. Right panel: CTA. Standard errors are displayed.

considering the unknowns had no significant effect, −3.9% versus −2.6%, $b = 1.3$, 95% CI $[-4.0, 6.2]$, $p > 0.5$. The interaction between domain type and treatment was marginally significant, indicating that the effect of considering the unknowns was larger in overconfident domains, with a 8.5% reduction in confidence after considering unknowns in overconfident domains compared to a 1.3% reduction in calibrated/underconfident domains, $b = 7.2$, 95% CI $[-0.4, 14.8]$, $p = 0.063$. In the CTA condition, neither of the simple effects was significant and there was no significant interaction (all p-values > 0.5).

Discussion

As in Study 2, considering the unknowns reduced confidence and overconfidence, but did not affect percent correct. The robustness of these effects to 2AFC versus 4AFC, within versus between participants, and with randomly versus nonrandomly sampled questions suggests that considering the unknowns is an effective debiasing technique under a variety of conditions. Importantly, considering the unknowns selectively reduced confidence in domains where participants were overconfident. We found some evidence that considering the alternative has some efficacy at reducing overconfidence (consistent with Koriat et al. 1980), but the effect of this manipulation was not consistent across our studies. We found some increase in percent correct in Study 2 but no effect on confidence and a small effect on confidence in Study 3 but no effect on percent correct. Across the two studies, considering the unknowns was more effective than considering the alternative at reducing confidence and equal to or better at reducing overconfidence.

General Discussion

Our studies show that the evaluation of what evidence is unknown or missing is an important determinant of judged confidence. However, people tend to underappreciate what they don't know. Thus, overconfidence is driven in part by insufficient consideration of unknown evidence.

We conceptualize known unknowns as evidence relevant to a probability assessment that a judge is aware that he or she is missing while making the assessment. We distinguish this from unknown unknowns—evidence that a judge is not aware he or she is missing. It is useful at this point to further distinguish two varieties of unknown unknowns. In some cases, a judge may be unaware that he or she is missing evidence but could potentially recognize that this evidence is missing if prompted. We refer to these as retrievable unknowns. In other cases, a judge is unaware that he or she is missing evidence and furthermore would need to be educated about the relevance of that evidence in order to recognize it as missing. We refer to these as irretrievable unknowns. To illustrate the importance of these distinctions, consider again the assessment of how likely it is that Iraq possesses nuclear weapons. In making this judgment, an intelligence analyst may explicitly ask herself whether Iraq possesses enriched uranium. The analyst may recall that enriched uranium is an important requirement for nuclear weapons, and that this factor is unknown. In this case, the question of whether or not Iraq has enriched uranium would be a known unknown. Alternatively, it may be that the analyst understands the relevance of uranium enrichment but does not consider this factor when judging the possibility of nuclear weapons. In this case, the presence of enriched uranium is a retrievable unknown. Studies 2 and 3 demonstrate the effectiveness of using a prompt to direct attention to retrievable unknowns that people may not otherwise consider, as a means of reducing misplaced confidence and improving calibration. However, consider further a nonexpert who does not know that enriched uranium is an important ingredient in nuclear weapons. In this case the presence of enriched uranium is an irretrievable unknown that a

CTU prompt could never elicit, though presumably the novice could be educated. This analysis predicts that a CTU prompt will only be effective in reducing misplaced confidence to the extent that the judge has sufficient expertise to recognize unknowns when prompted to do so.[3]

Our results suggest a potent new method that could be disseminated to practitioners for reducing overconfidence. First, "considering the unknowns" could be a self-administered treatment before making important judgments in situations where overconfidence is prevalent, such as when a CEO is making an acquisition (Malmendier and Tate 2005), when a CFO is budgeting for an upcoming year (Ben-David et al. 2007), or when a head of state is considering a military action (Johnson 2004).

Considering the unknown may also be a more effective debiasing technique than considering the alternative in some situations. In Studies 2 and 3, we compared CTU to CTA and found that considering the unknowns was more successful in reducing overconfidence. Further, we have provided some evidence that considering the unknowns selectively reduces confidence only when people are overconfident, whereas there is no evidence to suggest that reductions in confidence are selective when considering the alternative. Considering the unknowns may also be more effective than considering the alternative in judgment tasks where no obvious alternative exists. For instance, when estimating quantities in confidence intervals, such as "the cost of an advertising campaign," an instruction to "consider the alternative(s)" does not make sense (Alpert and Raiffa 1982). However, it may be possible to reduce interval overconfidence in such cases by prompting judges to consider the unknowns. This may be a fruitful area of future study because overconfidence is pervasive in confidence intervals estimation, with few techniques available to fully eliminate overconfidence biases (Klayman et al. 1999, Moore and Healy 2008, Soll and Klayman 2004).

Although we tout the potential of implementing a CTU strategy for debiasing, we do not claim that it will always outperform considering the alternative. One reason is that CTA can sometimes not only lead to reductions in confidence but also improvements in the proportion of items answered correctly (as we saw in Study 2). Additionally, considering the alternative may be a more viable approach when trying to de-bias others since it may be more compelling to argue for a concrete alternative option (playing "devil's advocate") than to argue that the other person is missing information (given that another person's retrievable unknowns are not necessarily retrievable to the persuader). Of course, these strategies are not mutually exclusive, and a hybrid strategy of considering both the *unknowns* and the *alternative* may be more effective than either strategy alone.

Donald Rumsfeld, secretary of defense during the invasion of Iraq in 2003 is famous for distinguishing between known knowns, known unknowns, and unknown unknowns. Our research suggests that the administration's overconfidence that Saddam Hussein possessed WMDs may have been due, in part, to focusing too much on the known knowns and neglecting the known unknowns. When Colin Powell made a speech to the UN Security Council in February of 2003 in which he presented a persuasive series of known facts supporting the existence of WMDs in Iraq he stated, "My colleagues, every statement I make today is backed up by sources, solid sources. These are not assertions. What we're giving you are facts and conclusions based on solid intelligence." If Colin Powell wanted his audience to have a more balanced view, he should have also articulated what was unknown to the Bush administration. Known unknowns could have ultimately strengthened or weakened the case for WMDs once they were resolved. For example, U.S. officials might have explicitly acknowledged how little they understood about Mr. Hussein's possible motivations for remaining coy about his nuclear program and moderated their confidence. Recently, it has come to light that Mr. Hussein was far more concerned about an internal coup or a Shiite rebellion than he was about a U.S. invasion, and so he encouraged everyone—from opponents in Iran to his own generals—to believe that he might have WMDs (Gordon and Trainor 2006). Our studies suggest there would have been little downside to U.S. officials considering what is unknown, at least from a judgment perspective. If unknowns are high, considering the unknown just might reduce overconfidence; and if unknowns are low, considering unknown evidence will not impact calibration.

Acknowledgments

D. J. Walters current affiliation is Marketing Area, INSEAD Europe Campus, Boulevard de Constance, 77305 Fontainebleau, France.

Endnotes

[1] Letter to President Bush, signed by Senator Bob Graham and others, December 5, 2001.

[2] Senator Jay Rockefeller, October 10, 2002.

[3] Evidence that is recognized to be unknown may also vary in terms of its specificity. For example, when predicting the outcome of a football game, a judge might consider that the health of the starting quarterback for the home team has been in question and so it is unknown whether the backup will have to carry the offense—a specific known unknown. Alternatively, a judge might consider that the variables that determine how the respective offenses and defenses of the teams match up is beyond his or her knowledge—a general known unknown. When debiasing using our "considering the unknowns" prompt, it is not clear to us which class of unknowns, the specific or the general, will tend to have a stronger effect on confidence and calibration.

References

Alpert M, Raiffa H (1982) A progress report on the training of probability assessors. Kahneman D, Slovic P, Tversky A, eds. *Judgement Under Uncertainty: Heuristics and Biases* (Cambridge University Press, New York).

Alter AL, Oppenheimer DM, Zemla JC (2010) Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *J. Personality Soc. Psych.* 99(3):436–451.

Babad E (1987) Wishful thinking and objectivity among sports fans. *Soc. Behaviour* 2(4):231–240.

Baumann AO, Deber RB, Thompson GG (1991) Overconfidence among physicians and nurses: The "micro-certainty, macro-uncertainty" phenomenon. *Soc. Sci. Medicine* 32(2):167–174.

Ben-David I, Graham JR, Harvey CR (2007) Managerial overconfidence and corporate policies. NBER Working Paper 13711, National Bureau Economonic Research, Cambridge, MA.

Brenner L, Griffin D, Koehler DJ (2005) Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organ. Behav. Human Decision Processes* 97(1):64–81.

Buehler R, Griffin D, Ross M (1994) Exploring the "planning fallacy": Why people underestimate their task completion times. *J. Personality Soc. Psych.* 67(3):366–381.

Camerer C, Lovallo D (1999) Overconfidence and excess entry: An experimental approach. *Amer. Econom. Rev.* 89(1)306–318.

Erev I, Wallsten TS, Budescu DV (1994) Simultaneous over- and underconfidence: The role of error in judgment processes. *Psych. Rev.* 101(3):519–527.

Feiler DC, Tong JD, Larrick RP (2013) Biased judgment in censored environments. *Management Sci.* 59(3):573–591.

Fernbach PM, Rogers T, Fox CR, Sloman SA (2013) Political extremism is supported by an illusion of understanding. *Psych. Sci.* 24(6):939–946.

Ferrell WR, McGoey PJ (1980) A model of calibration for subjective probabilities. *Organ. Behav. Human Performance* 26(1):32–53.

Fischhoff B, Slovic P, Lichtenstein S (1978) Fault trees: Sensitivity of estimated failure probabilities to problem representation. *J. Experiment. Psych.: Human Perception Performance* 4(2):330–344.

Fox CR (1999) Strength of evidence, judged probability, and choice under uncertainty. *Cognitive Psych.* 38(1):167–189.

Fox CR, Clemen RT (2005) Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Sci.* 51(9):1417–1432.

Gigerenzer G, Hoffrage U, Kleinbölting H (1991) Probabilistic mental models: A Brunswikian theory of confidence. *Psych. Rev.* 98(4):506–528.

Gordon MR, Trainor BE (2006) *Cobra II: The Inside Story of the Invasion and Occupation of Iraq* (Vintage, New York).

Griffin D, Brenner L (2004) Perspectives on probability judgment calibration. *Blackwell Handbook of Judgment and Decision Making*, 177–199.

Griffin D, Tversky A (1992) The weighing of evidence and the determinants of confidence. *Cognitive Psych.* 24(3):411–435.

Hoch SJ (1985) Counterfactual reasoning and accuracy in predicting personal events. *J. Experiment. Psych.: Learn., Memory, Cognition* 11(4):719–731.

Johnson DD (2004) *Overconfidence and War* (Harvard University Press, Cambridge, MA).

Johnson EJ, Häubl G, Keinan A (2007) Aspects of endowment: A query theory of value construction. *J. Experiment. Psych. Learn., Memory, Cognition* 33(3):461–474.

Kahneman D (2011) *Thinking, Fast and Slow* (Macmillan, New York).

Kahneman D, Tversky A (1973) On the psychology of prediction. *Psych. Rev.* 80(4):237–251.

Klayman J (1995) Varieties of confirmation bias. *Psych. Learn. Motivation* 32:385–418.

Klayman J, Ha YW (1987) Confirmation, disconfirmation, and information in hypothesis testing. *Psych. Rev.* 94(2):211–228.

Klayman J, Soll JB, Gonzalez-Vallejo C, Barlas S (1999) Overconfidence: It depends on how, what, and whom you ask. *Organ. Behavior Human Decision Processes* 79(3):216–247.

Koehler DJ, Brenner L, Griffin D (2002) The calibration of expert judgment: Heuristics and biases beyond the laboratory. Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge University Press, Cambridge, UK), 686–715.

Koriat A (2012) The self-consistency model of subjective confidence. *Psych. Review* 119(1):80–113.

Koriat A, Lichtenstein S, Fischhoff B (1980) Reasons for confidence. *J. Experiment. Psych.: Human Learn. Memory* 6(2):107–118.

Kunda Z (1990) The case for motivated reasoning. *Psych. Bull.* 108(3):480–498.

Larrick RP (1993) Motivational factors in decision theories: The role of self-protection. *Psych. Bull.* 113(3):440–450.

Mahajan J (1992) The overconfidence effect in marketing management predictions. *J. Marketing Res.* 29(3):329–342.

Malmendier U, Tate G (2005) CEO overconfidence and corporate investment. *J. Finance* 60(6):2661–2700.

McClelland AG, Bolger F (1994) The calibration of subjective probability: Theories and models 1980–94. Wright G, Ayton P, eds. *Subjective Probability* (John Wiley & Sons, Oxford, UK), 453–482.

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

Mynatt CR, Doherty ME, Tweney RD (1977) Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quart. J. Experiment. Psych.* 29(1):85–95.

Nickerson RS (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. General Psych.* 2(2):175–220.

Oskamp S (1965) Overconfidence in case-study judgments. *J. Consulting Psych.* 29(3):261–318

Rottenstreich Y, Tversky A (1997) Unpacking, repacking, and anchoring: Advances in support theory. *Psych. Rev.* 104(2):406–318.

Rozenblit L, Keil F (2002) The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Sci.* 26(5):521–562.

Sanbonmatsu DM, Kardes FR, Herr PM (1992) The role of prior knowledge and missing information in multiattribute evaluation. *Organ. Behav. Human Decision Processes* 51(1):76–91.

Sanbonmatsu DM, Kardes FR, Sansone C (1991) Remembering less and inferring more: Effects of time of judgment on inferences about unknown attributes. *J. Personality Soc. Psych.* 61(4): 546–554.

Schwenk CR, Cosier RA (1980) Effects of the expert, devil's advocate, and dialectical inquiry methods on prediction performance. *Organ. Behav. Human Performance* 26(3):409–424.

See KE, Fox CR, Rottenstreich YS (2006) Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. *J. Experiment. Psych.: Learn., Memory, Cognition* 32(6):1385–1402.

Soll JB (1996) Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organ. Behav. Human Decision Processes* 65(2):117–137.

Soll JB, Klayman J (2004) Overconfidence in interval estimates. *J. Experiment. Psych.: Learn., Memory, Cognition* 30(2):299–314.

Spiller SA, Fitzsimons GJ, Lynch JG Jr, McClelland GH (2013) Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *J. Marketing Res.* 50(2):277–288.

Tannenbaum D, Fox CR, Ülkümen G (2017) Judgment extremity and accuracy under epistemic vs. aleatory uncertainty. *Management Sci.* 63(2):497–518.

Tversky A, Kahneman D (1971) Belief in the law of small numbers. *Psych. Bull.* 76(2):105–110.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Tversky A, Koehler DJ (1994) Support theory: A nonextensional representation of subjective probability. *Psych. Rev.* 101(4):547–567.