Better Calibration When Predicting from Experience (Rather Than Description)

Forthcoming, *Organizational Behavior and Human Decision Processes*

Author Affiliations

Adrian R. Camilleri
Senior Lecturer
UTS Business School
University of Technology Sydney
14-28 Ultimo Rd, Ultimo NSW, 2007
Australia
E: adrian.camilleri@uts.edu.au

Ben R. Newell
Professor
School of Psychology
The University of New South Wales
Sydney, NSW, 2052
Australia
E: ben.newell@unsw.edu.au

Author's note

AC co-designed the studies, managed the data collection process, analyzed the data, and wrote the first draft of the manuscript. BN co-designed the studies and contributed to the manuscript writing. The authors thank Don Moore, Jack Soll, Craig Fox, members of the RMIT Behavioural Business Lab, and members of the UNSW Cognition Lab for helpful comments on this project. The author also thanks Mirra Seigerman for valuable research assistance. Experiment 3 was pre-registered at https://osf.io/4dcs8/

Abstract

The over-precision bias refers to the tendency for individuals to believe that their predictions are much more accurate than they really are. We investigated whether this type of overconfidence is moderated by how task-relevant information is obtained. We contrast cases in which individuals were presented with information about two options with equal average performance – one with low variance the other with high variance – in experience format (i.e., observed individual performance outcomes sequentially) or description format (i.e., presented with a summary of the outcome distribution). Across three experiments, we found that those learning from description tended to be over-precise whereas those learning from experience were under-precise. These differences were driven by a relatively better calibrated representation of the underlying outcome distribution by those presented with experience-based information. We argue that those presented with experience-based information have better learning due to more opportunities for prediction-error.

Keywords

Overconfidence, over-precision bias, risky choice, probability estimates, description-experience gap

Better Calibration When Predicting from Experience (Rather Than Description)

Robert Nardelli became CEO of Home Depot in December 2000 after an exhaustive external search by the board of directors (Lublin, Murray, & Brooks, 2000). Nardelli dramatically overhauled the company and replaced its entrepreneurial culture of innovative product design with one focused on relentless cost-cutting. During Nardelli's seven-year tenure, Home Depot stock remained stable while its competitor, Lowe's, stock doubled (Ries, 2007). Subsequently, Condé Nast Portfolio named Nardelli as one of the "Worst American CEOs of All Time" (CNBC, 2009). After Nardelli resigned as CEO on January 3, 2007, the Home Depot board promoted Frank Blake, who had worked diligently at the company for 5 years. During Blake's seven-year tenure as CEO, the company consistently outperformed Lowe's with stock rising by more than 145% and customer satisfaction increasing steadily each year (Aluise, 2012).

One of the key differences between the hires of Nardelli and Blake is that the former was brought in to be CEO whereas the latter was promoted to CEO. Although many factors likely determined the board's CEO decision, the one we focus on here is the way the board learned about the capabilities of these two men. Nardelli was learned about, in a large part, based on the strength of his achievements as described in his curriculum vitæ and by referees. In contrast, Blake was learned about, in a large part, based on the strength of his achievements as directly experienced by the board during his time as executive vice president of Home Depot. Could this difference in learning format – what we will refer to as description versus experience – have contributed to the quality of the board's decision and their confidence in it? In trying to answer this question, we draw on two phenomena from the prediction and risky choice literature: the over-precision bias and the description-experience gap.

A common finding in the prediction literature is a type of overconfidence called the "over-precision bias": excessive certainty regarding the accuracy of one's judgments (Moore & Healy, 2008). Over-precision could manifest as the Home Depot board over-estimating the positive effect of their new CEO hire on future stock prices. Such over-precision has been implicated in a range of real-world situations

including trading decisions (Odean, 1998), insurance purchase decisions (Silver, 2012), and advice taking (Yaniv, 2004).

A common finding in the risky choice literature is the "description-experience gap": choices made between risky options often vary depending on the format by which choice-relevant information is presented (Hertwig, Barron, Weber, & Erev, 2004; Hertwig, Hogarth, & Lejarraga, 2018). The two formats most frequently contrasted in this literature are descriptions and experiences. Experience formats involve the *sequential* presentation of outcome information in the form of *individual* statements (e.g., a daily summary report over a two-week period read each day by a front-line manager). In contrast, descriptive formats involve the *simultaneous* presentation of outcome and probability information in the form of *summary* statements (e.g., a quarterly summary report read by an executive-level manager). The bulk of this research has found that equivalent risky choice scenarios often produce divergent preferences such that those who learn from description appear to choose as if more strongly overweighting low probability outcomes (Hertwig & Erev, 2009). Such a description-experience choice gap could manifest as a front-line and executive-level manager disagreeing about which worker is the better performer. In the last decade, the description-experience distinction has sparked broad insights into a range of related areas including investment risk appetite (Kaufmann, Weber, & Haisley, 2013), responses to climate change (Newell, Rakow, Yechiam, & Sambur, 2016; Weber, 2006), consumers use of online review scores (Camilleri, 2017; Wulff, Hills, & Hertwig, 2014), doctor-patient interactions (Li, Rakow, & Newell, 2009), and responses to terrorist threats (Yechiam, Barron, & Erev, 2005).

The question posed and answered in this paper is a simple one: Does the format by which information is presented – description or experience – influence prediction precision and, if so, why? The answer to this question is important because decision-makers often have an option regarding the format with which to receive information, and thus a clear answer to our question could be used to strategically tailor information. Even in contexts where information format cannot be tailored, an answer to our question could provide needed insight regarding where to direct de-biasing interventions.

The Over-precision Bias

Overconfidence has been defined in numerous ways (Moore & Healy, 2008). The type of overconfidence focused on here is called over-precision and is supported by a large literature demonstrating that people have much more confidence in the accuracy of their beliefs than those beliefs warrant (Moore, Tenney, & Haran, 2015). In a typical demonstration, individuals are asked to estimate some uncertain outcome – such as the high temperature in Sydney on the first day of next summer – by constructing a confidence interval around it. For example, an 80% confidence interval is constructed such that the person is 80% sure that the true value falls between two interval limits. In most instances, the hit rate of these predictions – that is, the percentage of intervals that include the true outcome – was less than the assigned confidence level (Moore & Healy, 2008). For example, Soll and Klayman (2004) asked participants to construct fifty confidence intervals across a range of domains. Overall, participants' 80% intervals contained the correct answer only 48% of the time.

Several theories have been proposed to explain the over-precision bias. According to an anchoring account, people provide confidence intervals that are too close to the best estimate (Tversky & Kahneman, 1974). According to a conversational norms account, people prefer to be informative at the expense of being accurate (Yaniv & Foster, 1995). According to a naïve intuitive statistician account, people make estimates based on a small sample and that sample often underestimates the variance in the population (Juslin, Winman, & Hansson, 2007). However, none of these theories has satisfactorily explained all of the observed phenomena (Moore, Tenney, et al., 2015).

A notable feature of typical over-precision designs is that no learning takes place during the study. Rather, predictions are based on pre-existing knowledge brought into the study that was originally acquired for other purposes and in other contexts. For example, reflect on the knowledge you would bring to bear when making a prediction about Sydney's temperature. We believe this approach hinders theoretical progress because each person has different knowledge, and thus the source of the over-precision bias cannot be isolated. For example, the bias may derive from inaccurate knowledge, inappropriate extrapolation from existing knowledge, or both. There are two notable exceptions.

First, Goldstein and Rothschild (2014) examined over-precision in the context of "experience-based" knowledge. Participants learned about 100 outcomes that were sequentially presented. Participants were later asked to estimate their confidence in predicting the outcome of a new sample from the same distribution. Estimates were elicited by one of several stated formats (e.g., 80% confidence interval, fractiles) or by asking the participants to build a complete outcome distribution using a graphical interface (similar to the SPIES method; see Haran, Moore, & Morewedge, 2010). Note that by using this design the participant's prediction-relevant knowledge was known. Therefore, the researchers could calculate the accuracy of predictions in terms of the difference between the participant's prediction and the prediction of a rational agent with perfect memory. The researchers observed that, after learning from experience, predictions elicited via the graphical method were precise. For example, the 80% intervals derived from the graphical interface contained the correct answer 82% of the time. This is one of the few occasions in which over-precision has not been observed and suggests the intriguing possibility that experience-based learning may have a different effect on precision judgments. However, over-precision was obtained when measured using other elicitation methods. For example, the 80% intervals derived from the stated fractile method contained the correct answer just 48% of the time.

Second, Moore, Carter, and Yang (2015; Study 1) examined over-precision in the context of "description-based" knowledge. Participants were asked to provide estimates to questions that completely specified the outcomes and their probabilities. For example, a lottery question read: "Suppose you are planning to participate in a lottery game. Each day there is a 60% chance you will win $1 and a 40% chance that you will lose $1. How much money will you end up with after 500 days?" Note again how the participant's prediction-relevant knowledge was known and therefore the researchers could calculate the accuracy of predictions. The researchers observed that, after learning from description, predictions were inaccurate. For example, the 90% confidence intervals contained the correct answer 70% of the time. Interestingly, this over-precision was due to the intervals being centered on the incorrect value rather than being too narrow. In fact, the intervals were, on average, much wider than necessary.

Our literature review highlights several lessons. First, task-relevant knowledge can be acquired in two different ways - experience and description - and the degree of over-precision may vary as a function of this information format. However, no study has yet directly compared these formats in terms of over-precision. Second, the size of the over-precision bias is moderated by the elicitation method. This suggests that it may be prudent to use multiple elicitation methods although there are good reasons to prefer methods that elicit the complete outcome distribution (see Haran et al., 2010, for a discussion). Third, it is theoretically useful to study over-precision in contexts where task-relevant knowledge is controlled so that knowledge differences can be ruled out, and predictions can be compared to optimal responses.

Fourth, there are a number of different ways that over-precision can be operationalized. The most common is the group-level percentage of confidence intervals that end up including the realized outcome, which we call the *hit rate*. Confidence intervals can be called "over-precise" if the hit rate is less than the targeted confidence interval. Another indication of over-precision is the *interval width*. An interval can be called "over-precise" if it is narrower than is warranted relative to the true outcome distribution (i.e., more precise than the empirical or true distribution). Yet another indicator is the mean absolute difference between the estimated and true probability of each possible outcome, which we call *calibration*. The smaller the mean absolute difference the better calibrated. Finally, and perhaps most confusingly, the different ways that over-precision can be operationalized can conflict. For example, Moore, Carter, et al. (2015) observed that estimates were both over-precise (in terms of hit rate) but also under-precise (in terms of interval width). This is possible because the hit rate is evaluated relative to the obtained (or mean expected) outcome whereas interval width is evaluated relative to the true outcome distribution. We believe that interval width is not a good measure of prediction precision. This is because a person can perfectly match the variance and shape of the true outcome distribution – thus classifying as "precise" in terms of interval width – and yet be entirely wrong with all predictions because the distribution is misplaced (refer to Appendix A for further discussion). Although hit rate can also be problematic (because an interval of infinite width may achieve a perfect hit rate at the expense of any practical value),

it is the most common operationalization of precision and thus useful to connect with existing literature. Therefore, we decided to limit our analysis to (1) hit rate, which would coarsely define whether a group was over- precise, precise, or under-precise, and (2) calibration, which would define how closely the provided outcome distribution matched the true outcome distribution.

The Description-Experience Gap

In the context of risky choices, it has been argued that a continuum exists with regards to how uncertainty information is presented (Camilleri & Newell, 2013). At one end of the continuum is "experience-based" information in which individual outcomes are presented sequentially and their probabilities can only be inferred (e.g., Goldstein & Rothschild, 2014). At the other end of the continuum is "description-based" information in which outcomes and their probabilities are explicitly specified in a summary (e.g., Moore, Carter, et al., 2015).

The major observation from this stream of research is that people tend to make choices as if giving less weight to rare outcomes that are experienced versus those that are described (Hertwig & Erev, 2009; Rakow & Newell, 2010). In a typical demonstration of this "description-experience gap", participants are presented with a binary choice between a safe option, which pays one outcome with certainty (e.g., $3), and a risky option, which pays one of two outcomes with some probability (e.g., $4 with 80% chance, otherwise $0). Participants in the description version of the task are presented with a summary statement outlining the outcome distributions associated with each option and then asked to choose a preferred option. Participants in the experience version of the task must sequentially sample individual outcomes from each option, in any order and as often as desired, and are then asked to choose a preferred option. The choice tasks can be considered equivalent in that the sampled outcomes by those in the experience group are randomly selected from the same distributions that are stated to those in the description group. Hertwig et al. (2004) found that 36% of people in the description group preferred the risky option (i.e., 80% $4) to the safe option (100% $3), whereas 88% of people in the experience group preferred the risky option to the safe option. More recent research suggests that this gap is robust across a range of different contexts and problems (Wulff, Canseco, & Hertwig, 2018). However, there are some contexts where there

is no gap (Glöckner, Hilbig, Henninger, & Fiedler, 2016), particularly situations in which a small sample of perfectly representative outcomes is observed (Camilleri & Newell, 2011).

A number of theories have been put forward to explain the description-experience gap (see Hertwig, 2012 for a review). According to one early account, the gap occurs because people systematically misrepresent options' outcome distributions (Fox & Hadar, 2006). For example, people may overestimate the probability of rare events when learning from description but underestimate them when learning from experience. In order to assess this theory, several studies have asked participants to provide subjective estimates of experienced outcome probabilities. These studies typically present participants with a list of potential outcomes and then ask them to explicitly state the probability of each outcome occurring. For example, in one study, participants were asked to fill in the sentence, "__% of cards were worth 4 points" (Gottlieb, Weiss, & Chapman, 2007). In general, when using these methods, people produce estimates that are well calibrated (Fox & Hadar, 2006) or that *overestimate* (not underestimate) rarely experienced events (Barron & Yechiam, 2009; Camilleri & Newell, 2009; Hau, Pleskac, Kiefer, & Hertwig, 2008; Hertwig, Pachur, & Kurzenhauser, 2005; Ungemach, Chater, & Stewart, 2009). Thus, there is little support for the notion that the gap is driven by systematically misrepresented outcome distributions[1].

Despite the burgeoning literature, existing observations shed little light on how description- and experience-based knowledge might moderate prediction precision for three reasons. First, most previous studies asked participants to estimate the outcome distribution that had been observed in the past rather than the outcome distribution predicted in the future. Second, most previous studies elicited estimates only for experience-based information and not description-based information. Third, most previous studies have prevented participants from indicating a belief in future outcomes occurring that were not

[1] It is worthwhile to clearly distinguish the terms "over-estimate" and "over-weight": "Over-estimate" is a term that relates to belief in the probability that an outcome will occur. If an outcome has a 20% chance of occurring and yet a person believes that it has a 30% chance of occurring, then we would say this person has over-estimated the 20% outcome. "Over-weight" is a term that relates to how much impact an outcome has when making a choice. If a risk-neutral person prefers "$3 with 100% probability" over "$4 with 80% probability, else $0", then we would say that this person has over-weighted the $0 outcome. There are many reasons a person might over-weight an outcome, and over-estimating its likelihood is just one of them.

part of the original objective outcome distribution. Therefore, in this literature, participants' true predicted outcome distribution has remained unknown.

Hypothesis Development

It has been argued that the brain can be thought of as a hypothesis-testing, prediction-machine (Friston, 2005; Hohwy, 2013). Indeed, a primary function of memory is to predict the future (see Schacter et al., 2012 for a review). Format dependent-differences in prediction and, more specifically, the precision of those predictions could occur at two stages. First, there could be differences in how (the equivalent) information is encoded and represented in the mind. Second, there could be differences in how that information is used to generate predictions. In this paper, we argue that differences observed in prediction (i.e., the second stage) begin with differences at encoding (i.e., the first stage).

According to several influential theories, learning occurs via prediction errors (Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972). A prediction error refers to the discrepancy between what occurs and what was predicted to occur (Den Ouden, Kok, & De Lange, 2012). Experimental research has confirmed that prediction errors produce surprise, which is crucial for learning (Kamin, 1969). Note that the experience format, where outcomes are presented sequentially, naturally allows for a surprise: The presentation of each outcome affords an opportunity to make an implicit prediction about the next outcome, experience a prediction error, and improve learning. Consistent with this idea, there is considerable evidence demonstrating that learning from experience leads to good appreciation for properties of the underlying distribution (Hasher & Zacks, 1979; Hasher & Zacks, 1984; Hogarth & Soyer, 2011; Kaufmann et al., 2013; Zacks & Hasher, 2002) though there are some environments where this is not the case (Hogarth, Lejarraga, & Soyer, 2015). Indeed, repeated performance feedback has been shown to improve prediction accuracy (Stone & Opel, 2000). It is, therefore, no surprise that many experience-based choice models incorporate a prediction-error component (see Erev et al., 2010). In contrast, the description format, which summarises the entire outcome distribution in a single statement, does not naturally allow for a surprise: There is no opportunity to make a prediction, experience a prediction error, and improve learning. Consistent with this idea, there is evidence that learning from

description can lead to poor appreciation for properties of the underlying distribution (Camilleri & Newell, 2009; Erev, Glozman, & Hertwig, 2008; Gottlieb et al., 2007; Hawkins, Hayes, Donkin, Pasqualino, & Newell, 2015; Hoffrage, Krauss, Martignon, & Gigerenzer, 2015; Hogarth & Soyer, 2011). Therefore, it seems reasonable to expect that people learn the properties of an outcome distribution better from experience than description.

If predictions are based on the stored representation of the outcome distribution, then a relatively good underlying representation of the outcome distribution when learning from experience should produce a number of measurable downstream consequences for predictions. For example, we would expect better overall calibration and a hit rate in line with the confidence interval target. Additionally, we were interested in people's prediction of extreme events. An extreme event is one that is distant from the mean. Extreme events also tend to be rare; that is, have historically low occurrence. We focus on extreme events because excessive belief in such events can have a large impact on behavior (Lichtenstein, Slovic, Fischhoff, & Combs, 1978). Consequently, much of the research in the risky choice, especially the description-experience gap literature, has focused on skewed distributions with rare, extreme events, which is where we begin our investigation, too. One consequence of the expected poor outcome distribution by those learning from description is more random predictions, which would manifest as a higher expectation of extreme events.

Our conceptual development leads to the following formal hypotheses:

H1: When asked to predict future outcomes:

A: average confidence interval hit rate will be higher for those presented with information in experience (vs. description) format.

B: average calibration will be better for those presented with information in experience (vs. description) format[2].

[2] Given that we did not disclose the causal mechanism underlying the outcome distribution, we proceeded under the assumption that it would be rational to predict a future outcome distribution that was equal to the past outcome distribution. In practice, we did not expect the average participant to simply predict exactly the presented outcomes. Rather, we expected participants to invent a plausible causal mechanism and then try to predict future outcomes based on the recalled information, the invented

C: fewer extreme events will be predicted by those presented with information in experience (vs. description) format.

H2: When asked to recall past outcomes:

A: average calibration will be better for those presented with information in experience (vs. description) format.

Finally, a core assumption of our conceptualization is that predictions and choice rely on the same underlying knowledge representations. This assumption is consistent with many current theories of choice (Kiani & Shadlen, 2009; Merkle & Van Zandt, 2006; Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2013; Van den Berg et al., 2016). Therefore, we were also interested to explore how well estimates of the underlying outcome distribution predicted choices.

The Experiments

To explore how the over-precision bias is moderated by the format in which information is presented, we conducted three experiments. In each experiment, we presented the participant with information regarding the performance of a consistent worker and an inconsistent worker with equal average outcomes over a 10-day period. We manipulated the format in which worker information was presented. Participants were asked to fire one worker and retain the other. Participants were then asked to make predictions about each worker's future performance. In Experiment 1, which tested H1A, we presented participants with information in the description or experience format, and then asked for explicit confidence intervals. In Experiment 2, which tested H1ABC and H2, we introduced new problems, and elicited the entire outcome probability distribution with respect to what had been observed in the past or what was expected to be observed in the future. In Experiment 3, which tested H1ABC, we examined a broader range of choice problems using an incentive-compatible design.

Our research makes a number of contributions to the literature. Substantively, we contribute by answering the question of how the over-precision bias is moderated by the way that information is

causal mechanism, and some kind of prediction mechanism (Gaissmaier & Schooler, 2008; West & Stanovich, 2003). We expand on this interpretation in the General Discussion.

learned. In short, we find that those who learn from description tend to be over-precise (i.e., the group confidence interval hit rate is less than the target) whereas those who learn from experience tend to be under-precise (i.e., the group confidence interval hit rate is more than the target). This is an important finding because, historically, over-precision has been very difficult to eliminate, let alone reverse (Moore, Tenney, et al., 2015). Additionally, we reveal that those presented with information in experience format tend to learn the underlying outcome distribution better than those presented with description. This non-intuitive observation suggests that precision differences begin with what is learned rather than how that information is used. Finally, we show a close connection between the predicted (vs. recalled) outcome distributions and choice preference, suggesting that both judgments and choice derive from the same underlying representations of the alternative options.

Methodologically, we contribute by designing a procedure that tightly controls the information that people have when making a prediction, thus overcoming a limitation of previous studies in which researchers were blind to each participant's prediction-relevant knowledge. Additionally, we develop a method to collect probability outcome distributions that does not rely on understanding probabilities, which is a barrier for many participants. Finally, we allow participants to express belief in outcomes that they have never previously observed, thus capturing a truer representation of participant's perceived outcome distribution than most previous research.

Theoretically, we contribute by discussing a unique account for our observations: better learning for those presented with the experience format due to more opportunities for prediction error. This is a unique mechanism rarely discussed in the over-confidence and description-experience gap literature that is nevertheless fundamental to human learning. Additionally, we sketch out an exemplar-based model that could be used to make additional hypotheses about judgment and choice behavior when learning from description and experience. Finally, we contribute to the description-experience risky choice literature by providing a novel explanation for the description-experience gap: a higher expectation for previously unobserved outcomes, which often manifests as a higher belief in rare, extreme outcomes.

Experiment 1

The purpose of Experiment 1 was to test H1A. To do this, we presented participants with information in the description format, experience format, or both. Consistent with the bulk of existing literature, we asked participants to explicitly generate a 90% confidence interval for the predicted future performance of each worker. We began our investigation with a single choice problem in which the inconsistent worker sometimes performed well below the mean (i.e., a rare, extreme, bad outcome).

Methods

Participants

We aimed to collect data until there were more than 200 participants (i.e., on average 50 participants per group). This was an intuitive stopping rule. The final sample of participants were 202 Americans and Canadians (117 female, M_{age} = 33.4) recruited from Amazon's Mechanical Turk in exchange for money. Final group sizes ranged between 49 and 51.

Materials and Procedure

Each participant was asked to take on the role of a front-line manager, compare the performance of two workers, and choose to keep only one. One worker had consistent (i.e., low variance) performance and the other worker had inconsistent (i.e., high variance) performance. The mean performance of each worker was the same. It was also stated that the maximum number of sales possible on any single day was 20. Before proceeding to the choice phase, the participant was required to correctly answer 3 comprehension questions associated with the instructions.

During the choice phase, the participant was presented with information regarding 10 days of work for each worker before having to select one worker to keep. Information about both workers always appeared on the same screen together. That is, information about the two options was presented simultaneously.

After choosing, participants indicated their level of confidence in their choice ("How confident are you that you have selected the best salesperson for the long-term?") on a scale of 1 ("Not confident at

all") to 10 ("Extremely confident")[3]. Next, participants indicated 90% confidence intervals around each workers' expected average future performance by specifying the lower and upper bounds (Appendix B). On the next page, we attempted to elicit an entire outcome distribution for each option[4]. Finally, demographic information was collected, and the participants were thanked and paid.

Design

The full design involved random allocation to one of four information groups; however, here we focus on only the two that have direct relevance to our hypotheses. For those in the experience group, worker performance was presented as individual outcomes one at a time, side-by-side for each worker (see Appendix C). Specifically, the low variance worker's performance was 8, 8, 8, 8, 8, 9, 9, 9, 9, 9 (*SD* = 0.5) and the high variance worker's performance was 1, 9, 9, 9, 9, 9, 9, 10, 10, 10 (*SD* = 2.7), which we will call Problem 1. Outcomes were presented in one of 10 pre-determined orders that systematically varied where the rare outcome appeared in the sequence. Note that the average number of sales was 8.5 sales for both workers. For those in the description group, worker performance was described in a summary sentence, side-by-side for each worker (see Appendix D)[5]. There was also an option location variable that determined whether the low or high variance option was positioned on the left or right of the screen. The two prediction-related dependent variables were hit rate and choice. To calculate the hit rate, we coded whether ("1") or not ("0") the elicited 90% confidence interval included the true mean (i.e., the outcome "8.5"). Additionally, choice was coded in terms of whether the high variance option was selected ("1") or not ("0").

Results

[3] There was no difference in stated confidence between groups, $F(4, 197) = 1.11$, $p = .35$. We do not discuss this measure any further.

[4] For each option, we asked participants to adjust 21 bars indicating the probability of each potential outcome. We decided to leave out analysis of this question in light of feedback from participants that this tool was confusing to use. In particular, many participants struggled with the fixed sum nature of the tool. Another pilot study using only this 21 bars measure (i.e., no explicit confidence interval) was also conducted, which we also do not report on for the same reason. In Experiment 2, we designed a more intuitive question to elicit the entire outcome distribution.

[5] The additional groups comprised participants who were given both description- and experience-based information (either description then experience or vice versa). For ease of exposition and because these groups do not bear directly on the key hypotheses under test, we do not consider them in the main manuscript. However, full details can be found in Appendix E.

We analyzed the data using logistical regression with *format* as the independent variable. We also entered the counterbalancing variable, *option location*, as a covariate. For all experiments, we report all effects that passed the significance threshold ($p < .05$). To maintain clarity, in general, we report in footnotes any significant effects that are unrelated to the hypotheses.

Hit Rate

The hit rate of the 90% confidence intervals for each option is presented in Figure 1. Overall, the results support H1A. For the low variance option, the *hit rate* was significantly higher for those in the experience (vs. description) group, $\chi^2(N = 100) = 6.42$, $p = .01$. Follow-up analyses revealed that the *hit rate* for those in the experience group was not significantly different from 90%, $\chi^2(N = 51) = 0.28$, $p = .60$, whereas the *hit rate* for those in the description group was significantly less than 90%, $\chi^2(N = 49) = 10.76$, $p = .001$.

Similarly, for the high variance option, the *hit rate* was significantly higher for those in the experience (vs. description) group, $\chi^2(N = 100) = 10.94$, $p = .0009$. Follow-up analyses revealed that the *hit rate* for those in the experience group was not significantly different from 90%, $\chi^2(N = 51) = 2.66$, $p = .10$, whereas the *hit rate* for those in the description group was significantly less than 90%, $\chi^2(N = 49) = 10.76$, $p = .001$.

Choice

The proportion of high variance choices was higher for those in the experience group (.65) than those in the description group (.35), $\chi^2(N = 100) = 9.22$, $p = .002$.

To investigate the connection between choice and estimates, we computed a variable predicting whether the low or high variance option was expected to be selected based on the participant's relative confidence intervals. Specifically, we subtracted the implied mid-point of the high variance confidence interval from the implied mid-point of the low variance confidence. Based on this exercise, 43 people (67% in the description group) were predicted to choose the low variance option, 36 people (33% in the description group,) were predicted to select the high variance option, and 21 people (38% in the description group) had no prediction (because of equal midpoints). The proportion of choices correctly

predicted for cases when a prediction was possible was 0.87, which is significantly higher than chance, $\chi^2(N = 79) = 49.50$, $p < .0001$. There was no difference in the proportion of choices correctly predicted between formats, $\chi^2(N = 79) = 0.01$, $p = .92$.

Discussion

The results of Experiment 1 provide initial support for our conceptualization. Based on observed hit rates, H1A was supported: those who learned from description were *over-precise* (i.e., interval hit rate lower than the 90% target) whereas those who learned from experience were precise (and, if anything, trended towards being *under-precise* with an interval hit rate higher than the 90% target).

Although we made no explicit hypothesis regarding choice patterns, our observations replicated the classic description-experience choice gap: Those who learned from experience were more likely than those learning from description to choose the high variance option. This pattern is consistent with those learning from description overweighting and/or overestimating the likelihood of the rare event, which in this case was an extreme, "bad" outcome (i.e., much lower than the mean outcome). Interestingly, we found that choices could be predicted much better than chance when given the (inferred) average outcome from each option. This result reveals a tight connection between estimates and choice, though the exact nature of this relationship cannot be determined when relying on the very coarse hit rate variable. We explore this finding further in Experiment 2.

Experiment 2

The purpose of Experiment 2 was to test H1ABC and H2. To do this, we asked half of the participants to report on outcomes predicted in the future (as in Experiment 1) and asked the other half to report on outcomes observed in the past. We replaced the explicit interval question with a novel one that elicited the entire outcome distribution without mention of probabilities. This allowed us to test whether H1A was robust to different elicitation methods, and also permitted us to evaluate the calibration of the predicted outcome distribution relative to the true outcome distribution (i.e., mean absolute error). In Experiment 1, the description-based statement was probabilistic, which research shows can be difficult to understand (Gigerenzer & Hoffrage, 1995; Gottlieb et al., 2007). Therefore, in Experiment 2, we replaced

probabilistic information with frequency information. To improve generalizability, we designed two new choice problems: one in which the inconsistent worker sometimes performed well below the mean (i.e., a rare, extreme, bad outcome), and the other in which the inconsistent worker sometimes performed well above the mean (i.e., a rare, extreme, good outcome).

Methods

Participants

We aimed to collect data until there were more than 600 participants (i.e., on average 75 participants per group). This was an intuitive stopping rule and larger than Experiment 1 given the more complex design. The final sample of participants were 604 Americans and Canadians (311 female, M_{age} = 32.3) recruited from Amazon's Mechanical Turk in exchange for money. Final group sizes ranged between 71 and 80.

Materials and Procedure

The procedure was identical to Experiment 1 with the following change: after the choice phase, instead of providing explicit confidence intervals, the participant was required to generate an outcome distribution for each worker by stating 10 outcomes for each worker. Note that this approach allowed participants to enter values larger than 20 even though it was stated in the instructions that this was not possible in the scenario.

Design

The participants were randomly allocated to one of eight groups according to a 2 (Problem: 1 vs. 2) x 2 (Format: Description vs. experience) x 2 (Estimate type: Past vs. future estimation) between-subjects design.

Problem was manipulated via the composition of the ten outcomes associated with the high variance option. For Problem 1, the high variance option outcomes were: 1, 1, 10, 10, 11, 11, 11, 11, 12, 12 (*SD* = 4.3). Problem 1, therefore, featured a bad extreme outcome that was also rare (i.e. 1). For Problem 2, the high variance option outcomes were: 6, 6, 6, 6, 7, 7, 8, 8, 18, 18 (*SD* = 6.8). Problem 2, therefore, featured a good extreme outcome that was also rare (i.e., 18). For both problems, the low variance option

outcomes were: 8, 8, 8, 9, 9, 9, 9, 10, 10, 10 (*SD* = 0.8). Note that the average number of sales was 9.0 for all problems and all options.

For those allocated to the description information format, worker performance was described in a summary sentence, side-by-side for each worker (see Appendix F). Unlike Experiment 1, the sentence did not make reference to any probabilities. For those allocated to the experience information format, worker performance was presented as individual outcomes one at a time, side-by-side for each worker (see Appendix C). Outcomes were presented in one of 10 pre-determined orders that systematically varied where the extreme outcomes appeared in the sequence.

For those allocated to the past estimate type, the outcome distribution question stated: "*Think back: please recall as best you can how many sales [the name of the worker] achieved each particular day*". For those allocated to the future estimate type, the outcome distribution question stated: "*Think forward: please estimate as best you can how many sales [the name of the worker] will achieve each particular day*" (See Appendix G).

To calculate the implied 80% confidence intervals, we ordered the provided 10 outcomes from smallest to largest, and then eliminated the first and last outcomes. To calculate the hit rate, we coded whether ("1") or not ("0") this implied interval included the true mean (i.e., the outcome "9"). To evaluate the calibration of the provided outcome distributions, we examined the mean absolute error between the estimated likelihood of each outcome and the true likelihood of each outcome after arranging both in order from lowest to highest. For each option we also computed a variable capturing the average proportion of "extreme outcomes" in the estimated outcome distribution of the two options. Extreme outcomes were operationalized as less than "3" or more than "17". These values were chosen somewhat arbitrarily but are robust to a sensitivity analysis (see footnote 6).

Results

We analyzed the data using ANOVA and logistical regression with *format*, *problem*, *estimate type*, and their interactions all entered as independent variables. We also entered the counterbalancing variable, *option location*, as a covariate.

Hit Rate

The hit rate of the implied 80% confidence interval for each option is presented in Figure 2. As in Experiment 1, H1A was supported. When making a prediction about the future for the low variance option, the *hit rate* was significantly higher for those in the experience (vs. description) group, $\chi^2(N = 304) = 11.03$, $p = .0009$. Follow-up analyses revealed that the *hit rate* for those in the experience group was significantly higher than 80%, $\chi^2(N = 157) = 14.49$, $p = .0001$, whereas the *hit rate* for those in the description group was not significantly different from 80%, $\chi^2(N = 147) = 0.28$, $p = .60$.

Similarly, when making a prediction about the future for the high variance option, the *hit rate* was significantly higher for those in the experience (vs. description) group, $\chi^2(N = 304) = 11.42$, $p = .0007$. Follow-up analyses revealed that the *hit rate* for those in the experience group was significantly higher than 80%, $\chi^2(N = 157) = 4.74$, $p = .03$, and the *hit rate* for those in the description group was significantly lower than 80%, $\chi^2(N = 147) = 6.16$, $p = .01$.

Calibration

The average estimated likelihood of each outcome is presented in Figure 3. The observations support H2. When recalling past outcomes for the low variance option, *calibration* was significantly better for those in the experience (vs. description) group, $F(1, 295) = 18.24$, $p < .0001$. Similarly, when recalling past outcomes for the high variance option, *calibration* was significantly better for those in the experience (vs. description) group, $F(1, 295) = 24.94$, $p < .001$.

The observations also support H1B. When making predictions about the future for the low variance option, *calibration* was significantly better for those in the experience (vs. description) group, $F(1, 299) = 17.59$, $p < .0001$. Similarly, when making predictions about the future for the high variance option, *calibration* was significantly better for those in the experience (vs. description) group, $F(1, 299) = 20.39$, $p < .0001$.

The observations also support H1C. When making predictions about the future for the low variance option, the proportion of extreme outcomes was significantly lower for those in the experience (vs. description) group, $F(1, 299) = 19.18$, $p < .0001$. When making predictions about the future for the high

variance option, the proportion of extreme outcomes was significantly lower for those in the experience (vs. description) group, $F(1, 299) = 14.51$, $p = .0002$[6].

Choice

The proportion of choices made in each group is presented in Figure 4. Estimate type was ignored because at the point of choice this factor was not yet implemented. Significantly more high variance choices were made by those in the description (vs. experience) group $\chi^2(N = 604) = 9.78$, $p = .002$. However, this effect was qualified by a significant interaction, $\chi^2(N = 604) = 10.04$, $p = .002$, indicating that the difference was driven by Problem 2 in which the extreme outcome was good[7].

To investigate the connection between choice and estimates, we computed a variable predicting whether the low or high variance option was expected to be selected based on the participant's relative outcome distributions. Specifically, we subtracted the implied mean of the high variance outcome distribution from the implied mean of the low variance outcome distribution. Based on this exercise, 275 people (47% in the description group) were predicted to choose the low variance option, 244 people (42% in the description group,) were predicted to select the high variance option, and 85 people (71% in the description group) had no prediction (because of equal means). The proportion of choices correctly predicted for cases when a prediction was possible is shown in Figure 5. Prediction success was significantly higher when based on future (vs. past) estimates, $\chi^2(N = 519) = 23.54$, $p < .0001$, and there was no effect of format, $\chi^2(N = 519) = 0.42$, $p = .51$)[8].

Discussion

[6] There was also a significant effect of problem type, $F(1, 299) = 6.62$, $p = .01$, reflecting that the proportion of extreme outcomes was significantly lower for Problem 2 (vs. 1). Results when extreme outcomes was operationalised as less than "2" or more than "18": For the low variance option, description ($M = 0.8$) was significantly higher than experience ($M = 0.01$), $F(1, 299) = 13.79$, $p = .0003$. For the high variance option, description ($M = 0.12$) was significantly higher than experience ($M = 0.08$), $F(1, 299) = 7.31$, $p = .007$. Results when extreme outcomes operationalised as less than "4" or more than "16": For the low variance option: description ($M = 0.13$) significantly higher than experience ($M = 0.03$), $F(1, 299) = 19.75$, $p < .0001$. For the high variance option, description ($M = 0.25$) was significantly higher than experience ($M = 0.16$), $F(1, 299) = 18.84$, $p < .0001$.

[7] There was also a significant effect of *option location*, $\chi^2(N = 604) = 4.04$, $p = .04$, reflecting that more high variance choices were made when that option was positioned on the right hand side of the screen.

[8] The analysis also revealed a significant effect for problem type, $\chi^2(N = 519) = 11.27$, $p = .0008$, reflecting better calibration for problem 1 (vs. 2).

The results of Experiment 2 reinforce and extend those of Experiment 1. Supporting H1A, we observed that the average confidence interval hit rate was higher for those presented with information in experience (vs. description) format. Whereas those learning from experience tended to be under-precise (i.e., 80% intervals contained the mean expected outcome more than 80% of the time), those learning from description tended to be over-precise for the high variance option (i.e., 80% intervals contained the mean expected outcome less than 80% of the time) or precise for the low variance option (i.e., 80% intervals contained the mean expected outcome approximately 80% of the time). In this experiment, we were able to more clearly understand the source of this discrepancy by eliciting the entire outcome distribution for past and predicted outcomes. Supporting H1B, the outcome distribution of those who had learned from description were poorly calibrated compared to those who had learned from experience. In particular, those who learned from description were more likely to expect extreme outcomes. This observation suggests better encoding of information by those who learned from experience. Better recall of the presented outcome distribution translated into better future predictions.

We found a description-experience choice gap in one out of the two problems examined in Experiment 2. A clue to the occurrence of the gap in one problem but not the other can be found in future estimates summarised in Figure 3 Panel A. As can be seen, there is little difference between description- and experience-based predictions for outcomes 8, 9, and 10, in Problem 1 but a large difference for Problem 2. Therefore, the gap in Problem 2 may be driven by those in the description group overestimating the number of (low outcome) extreme events associated with the low variance option.

A final noteworthy result is that we were able to correctly predict over 86% of choices by simply comparing the means of the option's estimated future outcome distributions. This finding lends support to the idea that estimates and choice are based on the same underlying representations. It is also important to note that choice prediction was significantly worse – closer to 67% – when based on comparing the means of the option's recalled outcome distributions. This finding lends support to the idea that people use or transform their stored representation of the outcome distribution when making a prediction. Finally, as in Experiment 1, we found no effect of format on the proportion of choices correctly predicted.

This result suggests that, despite format-dependent differences in what was learned, those underlying representations are used in the same way to guide a choice. We return to these ideas in the General Discussion.

Experiment 3

The purpose of Experiment 3 was to again test the first three hypotheses while addressing some limitations associated with the first two experiments. First, the earlier experiments relied on a sample recruited from Amazon's Mechanical Turk. Second, the earlier experiments did not incentivize participants based on their probability estimates and choices. This design might threaten the validity of our conclusions if one information format was inherently more interesting. Such a concern would be eliminated if our earlier observations replicated in a context where both probability estimates and choices were incentivized. Third, the earlier experiments used choice problems in which the high variance option was associated with a skewed distribution where the rare outcome was also the extreme outcome and located well above or below the mean. These choice problems might threaten the validity of our conclusions because the opportunity to overestimate (vs. underestimate) extreme outcomes was not equal. For example, if those presented with description-based information naturally have a greater tendency to overestimate rare outcomes, then this group would always produce less well calibrated distributions when rare outcomes are also extreme outcomes. Such a concern would be eliminated if our earlier observations replicated in problems that had a symmetrical distribution that separated the rare and extreme outcomes, particularly if the rare outcome was also the mean outcome.

To address these concerns, Experiment 3 relied on a different sample group, was incentive-compatible, and used new choice problems. One of the most important new choice problems was structured as follows: 3 with a 40% chance, 9 with a 20% chance, and 17 with a 40% chance. Note that in this problem, the mean outcome is 9, the rare outcome is also 9, and the extreme outcomes are 3 and 17. Therefore, if those learning from description simply over-estimated the rare outcome, they would also be over-estimating the mean outcome, which should result in higher hit rate. In contrast to this possibility, we again predicted that those learning from experience would learn the underlying outcome distribution

better, which would be reflected in better calibration (i.e., H1B) and higher hit rates (i.e., H1A). Importantly, we were expecting these predictions to hold across all four different problems, thus predicting no interaction between format and problem.

The prediction for H1C – fewer extreme events predicted by those presented with information in experience (vs. description) format – is more nuanced. This prediction is based on the expectation of more random predictions by those in the description group, due to poor learning of the true outcome distribution. This prediction makes sense when the true outcome distribution is normally distributed or skewed because (random) predictions of extreme outcomes likely reduce calibration. However, when the true outcome distribution is symmetrical and contains many extreme outcomes, such as the problem outlined above, then random predictions of extreme outcomes may (inadvertently) improve calibration. Therefore, we expected an interaction between format and problem for the high variance option.

Methods

Participants

We aimed to collect data until there were more than 200 participants, which was chosen to obtain at least 90% power to detect a small-to-medium effect size. The final sample of participants were 232 Australians (174 female, M_{age} = 20.6) recruited from a public university undergraduate student pool in exchange for course credit. Participants also had the opportunity to earn cash contingent on outcome distribution judgments and choices made during the experiment, as well as a lottery conducted after the experiment. Final group sizes ranged between 114 and 118.

Materials and Procedure

The procedure was identical to Experiment 2 with the following changes: First, participants made four (rather than one) evaluations between pairs of workers. Each choice was described as corresponding to a different part of the country: north-east, south-east, north-west, and south-west. Second, we made the experiment incentive compatible by paying 1 out of 20 participants based on their choices and on the calibration of judgments about the outcome distribution (see Appendix H for instructions). The choice payment was operationalized as the average 10-day future performance of one randomly selected worker

the participant had chosen to keep. The calibration payment was operationalized as the summed absolute distance between the forecasted and past percentage of sales for the same worker. Third, the estimated sales values for each day were restricted to be between 0 and 20, which was consistent with the instructions stating that 20 was the maximum number of sales per day per worker. Fourth, participants made predictions only about the worker's future performance.

Design

The participants were randomly allocated to one of eight groups according to a 4 (Problem: 1 vs. 2 vs. 3 vs. 4) x 2 (Format: Description vs. Experience) mixed-subjects design. *Problem* was manipulated within subject and *format* was manipulated between-subjects. In addition, there were two counterbalance variables: *problem order* and *option location*. The *problem order* variable determined the order in which the four problems were presented according to a Latin square. The *option location* variable determined whether the low or high variance option was positioned on the left or right of the screen.

Problem was manipulated via the composition of the ten outcomes associated with the high variance option. For Problem 1, the high variance option outcomes were: 1, 1, 11, 11, 11, 11, 11, 11, 11, 11 (*SD* = 4.2). Problem 1, therefore, featured a bad extreme outcome (i.e., 1) that was also rare. For Problem 2, the high variance option outcomes were: 7, 7, 7, 7, 7, 7, 7, 7, 17, 17 (*SD* = 4.2). Problem 2, therefore, featured a good extreme outcome (i.e., 17) that was also rare. For Problem 3, the high variance option outcomes were: 1, 1, 1, 1, 9, 9, 17, 17, 17, 17 (*SD* = 7.5). Problem 3, therefore, featured both bad and good extreme outcomes that were not rare. For Problem 4, the high variance option outcomes were: 1, 1, 9, 9, 9, 9, 9, 9, 17, 17 (*SD* = 5.3). Problem 4, therefore, featured both bad and good extreme outcomes that were also both rare. For all four problems, the low variance option outcomes were 8, 8, 8, 9, 9, 9, 9, 10, 10, 10 (*SD* = 0.8). Note that the average number of sales was 9.0 for all problems and all options. *Format* was manipulated in the same way as in Experiment 2.

Results

Our analysis relied on a series of linear mixed-effects models (LMM) and generalized linear mixed-effects models (GLMM). We preferred the mixed-effects model because it flexibly enables the modeling

of correlated data—inherent to the nature of our design—without the violation of important regression assumptions (Demidenko, 2004). In all models, the participant *ID* was entered as a random effect. We entered *format* (coded 0 = description, 1 = experience), *problem* (coded 1, 2, 3, or 4), and their interaction as independent variables. We also entered the two counterbalancing variables – *problem order* and *option location* – as covariates.

Hit Rate

The hit rate of the implied 80% confidence interval for each option is presented in Figure 6. To analyze this variable, we entered *hit rate* (coded 0 = Fail, 1 = Success) as the dependent variable in the binary logistical regression GLMM. Overall, the results again support H1A. For the low variance option, the *hit rate* was significantly higher for those in the experience (vs. description) group, $F(1, 916) = 19.77$, $p < .001$, and there was no interaction between format and problem, $F(1, 916) = 0.68$, $p = .57$. Follow-up analyses revealed that the *hit rate* for those in the experience group was significantly higher than 80%, $\chi^2(N = 472) = 73.07$, $p < .0001$, whereas the *hit rate* for those in the description group was not significantly different from 80%, $\chi^2(N = 456) = 0.82$, $p = .37$.

For the high variance option, the *hit rate* was also significantly higher for those in the experience (vs. description) group, $F(1, 916) = 8.58$, $p = .003$, and there was no interaction between format and problem, $F(1, 916) = 2.43$, $p = .06$[9]. Follow-up analyses revealed that the *hit rate* for those in the experience group was significantly higher than 80%, $\chi^2(N = 472) = 41.87$, $p < .0001$, whereas the *hit rate* for those in the description group was not significantly different from 80%, $\chi^2(N = 456) = 1.78$, $p = .18$.

Calibration

The average estimated likelihood of each outcome is presented in Figure 7. To analyze this variable, we entered *calibration* as the dependent variable in the LMMs. Overall, the results support H1B. For the low variance option, *calibration* was significantly higher for those in the experience (vs. description)

[9] In addition, this analysis revealed a significant effect for *problem*, $F(1, 916) = 6.70$, $p < .001$, and *option location*, $F(1, 916) = 4.00$, $p = .046$. That is, for the high variance option, the hit rate was relatively higher for problems 3 and 4 (vs. problems 1 and 2), and when that option was located on the right of screen.

group, $F(1, 232) = 27.12$, $p < .001$, and there was no interaction between format and problem, $F(3, 696) = 2.34$, $p = .07$[10]. For the high variance option, *calibration* was significantly higher for those in the experience (vs. description) group, $F(1, 232) = 26.78$, $p < .001$, and there was no interaction between format and problem, $F(3, 696) = 0.97$, $p = .41$[11].

The observations also support H1C. When making predictions about the future for the low variance option, the proportion of extreme outcomes was significantly lower for those in the experience (vs. description) group, $F(1, 232) = 10.72$, $p = .001$. When making predictions about the future for the high variance option, the proportion of extreme outcomes was significantly influenced by format, $F(1, 232) = 4.58$, $p = .03$, problem, $F(1, 696) = 94.50$, $p < .001$, and their interaction, $F(1, 696) = 8.82$, $p < .001$. As expected, the proportion of extreme outcomes was significantly lower for those in the experience (vs. description) group for Problems 1 and 2 (associated with a skewed distribution) but were no different for Problems 3 and 4 (associated with a symmetrical distribution with extreme outcomes).

Choice

The average proportion of high variance choices made in each group is presented in Figure 8. To analyze this variable, we entered *choice* (coded 0 = Low variance option, 1 = High variance option) as the dependent variable in the binary logistical regression GLMM. The analysis revealed no effect of format, $F(1, 916) = 0.29$, $p = .59$, nor was there an interaction between format and problem, $F(1, 916) = 1.95$, $p = .12$.

To investigate the connection between choice and estimates, we computed a variable predicting whether the low or high variance option was expected to be selected based on the participant's relative outcome distributions. Based on this exercise, 394 people (47% in the description group) were predicted to choose the low variance option, 331 people (48% in the description group) were predicted to select the high variance option, and 203 people (55% in the description group) had no prediction. The proportion of

[10] In addition, this analysis revealed a significant effect for *option location*, $F(1, 232) = 7.01$, $p = .009$. That is, for the low variance option, calibration was significantly better when the low variance option was located on the right of screen.
[11] In addition, this analysis revealed a significant effect for *problem*, $F(3, 696) = 7.45$, $p < .001$. That is, for the high variance option, calibration was significantly better for problem 3 (vs. problems 1, 2, and 4).

choices correctly predicted for cases when a prediction was possible was 0.70, which is significantly higher than chance, $\chi^2(N = 725) = 375.97$, $p < .0001$. A final binary logistical regression GLMM analysis with proportion of choices correctly predicted as the dependent variable revealed no significant effects (all p's > .05) suggesting that the model predictions were equally good for all formats and problems.

Discussion

The results of Experiment 3 reinforce those of Experiments 1 and 2. Supporting H1A, we observed that the average confidence interval hit rate was higher for those presented with information in experience (vs. description) format. Whereas those learning from experience tended to be under-precise (i.e., 80% intervals contained the mean expected outcome more than 80% of the time), those learning from description tended to be precise (i.e., 80% intervals contained the mean expected outcome approximately 80% of the time). We note that the hit rate of those in the description group varied considerably across experiments and options. Inspection of the hit rates suggests that the type of distribution matters: the hit rate tended to be higher when the outcome distribution was symmetrical (for example, in Figure 6 Panel B, compare the non-symetrical Problems 1 and 2 with the symetrical Problems 3 and 4). This is because random predictions are more likely to decrease the hit rate in contexts with a skewed true outcome distribution. A clear implication is that skewed distributions are more diagnostic problems for evaluating prediction precision.

In support of H1B, the outcome distribution of those who had learned from experience was better calibrated than those who had learned from description. This suggests better encoding of information by those who learned from experience. Interestingly, calibration for the low variance option, which had the same distribution in Experiments 2 and 3, was better in Experiment 3 ($M = 1.34$, $SD = 1.87$) than Experiment 2 ($M = 1.67$, $SD = 2.51$), $t(1230) = 2.44$, $p = .01$. This difference is likely attributable to the effects of incentives, the different sample group, or both.

One question raised by these findings relates to how those learning from experience (vs. description) could be better calibrated and yet less precise. We believe this result again highlights how different operationalizations of precision provide different perspectives on exactly who is overconfident (see also

Appendix A). In these experiments, hit rate were scored relative to the expected mean outcome of the presented outcome distribution; in essence, all intervals were compared to a one-show draw and that draw was always the mean outcome. For example, for the high variance option of Problem 3, this was an outcome of 9. As a result, good calibration was associated with a very high hit rate for those learning from experience. An alternative way to score hit rates is relative to a random draw from the outcome distribution. For example, for the high variance option of Problem 3, 20% of the time this outcome would be 1, 20% of the time this outcome would be 17, and 60% of the time this outcome would be 9. We note that this second scoring rule would reduce overall hit rates and more closely align calibration and precision measures.

General Discussion

According to Moore, Tenney, et al. (2015), the form of overconfidence focused on in this paper – over-precision – is interesting because it is the "*… most robust form of overconfidence …* [with] *few, if any, documented reversals*". Much research has found that the truth is often surprisingly different from people's expectations (Alpert & Raiffa, 1982; Soll & Klayman, 2004; Yaniv & Foster, 1995). The current set of experiments is therefore particularly compelling because it documents one of the few examples of under-precision – situations in which the truth is quite similar to people's expectations.

In our studies, the degree of precision was measured in terms of hit rate: how often a confidence interval included the true expected mean relative to the assigned confidence level. Our key manipulation was, prior to judgment, to provide information about the underlying outcome distribution in one of two forms: as individual outcomes observed sequentially (i.e., by experience), or as a summary statement describing the distribution of outcomes (i.e., by description). The information provided was objectively equivalent yet the format clearly influenced precision levels: Those learning from experience tended to be under-precise whereas those learning from description tended to be over-precise or, sometimes, precise[12].

[12] It is worth highlighting that the overall hit rate in the current set of experiments is much higher than in prior research using traditional paradigms, where the hit rates rarely climb above 60% for 90% confidence intervals. One likely explanation for this is that all information relevant for making a prediction in our experiments was provided during the study itself, thereby reducing the impact of memory. A second factor is our unique elicitation method, which asked for 10-day forecasts, thus eliciting an outcome distribution without requiring the participant to explicitly communicate probabilities.

An individual who makes predictions that too often miss the mark may be constructing a confidence interval that is too narrow or constructing a confidence interval that is poorly calibrated with reality (or both). Our data suggest that the latter image is often the more fitting for those learning from description. Observations made in Experiment 2 revealed that recalled outcome distributions were much better calibrated when participants were presented with experience- compared to description-based information. These observations are consistent with findings demonstrating that individuals are good at retaining information acquired from the sequential presentation of outcomes (Goldstein & Rothschild, 2014; Hogarth & Soyer, 2011; Kaufmann et al., 2013; Peterson & Beach, 1967; Zacks & Hasher, 2002). In contrast, those presented with description-based information tended to estimate outcome distributions with a higher degree of noise, which was often reflected as over-predicting the likelihood of extreme outcomes. Therefore, our results suggest that at least part of the reason that those learning from description were over-precise was that their recollection of the initial task-relevant information was relatively poor.

Much past research has found a format-dependent difference in risky choice contexts consistent with the idea that rare outcomes are more overweighted in the description than experience format (see Wulff et al., 2018 for a review). We observed such a "choice gap" in Experiment 1 and also in Experiment 2 for one problem. However, we did not observe a gap for any of the four problems in Experiment 3. These observations are not unprecedented in light of studies finding small, zero, and sometimes even reversed choice gaps, particularly when people are presented with small representative samples such as in our experiments (Camilleri & Newell, 2011; Fox & Hadar, 2006; Glöckner et al., 2016).

How do these observations help us to understand past findings of overconfidence? First, let us reflect on the kind of questions participants have been presented with in previous studies. Typically, they have been general knowledge, trivia-type questions in domains ranging from science to history to sports. For example, Soll and Klayman (2004) asked participants to construct confidence intervals around the invoice price of a sedan, the winning percentage of a basketball team, colleges' overall quality score, average movie box office results, human fertility rates of different countries, the year in which a variety of devices

and processes were invented or discovered, and the average daily high July temperature of major cities around the world. For most of these questions, it is clear that knowledge was attained primarily from description. Take the question of estimating the high temperature in Sydney on the first day of next summer. Apart from those lucky enough to live in Sydney, all relevant knowledge must come from description (e.g., looking up the weather on a website; seeing the sunshine in a postcard photograph). Our contention is that description-based information is encoded in a fundamentally different way than experience-based information (Camilleri & Newell, 2013).

We propose that when learning from experience, each observed outcome is explicitly stored into memory to serve as the basis of a representation. This assumption is consistent with recent exemplar theories of choice (Ashby & Rakow, 2014; Gonzalez & Dutt, 2011; Hawkins, Camilleri, Heathcote, Newell, & Brown, 2014; Lejarraga, Dutt, & Gonzalez, 2012) and also many other areas of cognition including theories of categorization (Nosofsky, Palmeri, & McKinley, 1994). Correspondingly, we propose that when learning from description, people use the description to mentally simulate a set of sample outcomes that are explicitly stored into memory to serve as the basis of a representation. This assumption is also consistent with recent choice models attempting to capture the behavior of both description- and experience-based choices within a single framework (Erev, Ert, Plonsky, Cohen, & Cohen, 2017; Erev et al., 2008; Lin, Donkin, & Newell, 2015).

We propose that outcomes are imperfectly stored into memory (Hawkins et al., 2014; Lin et al., 2015). According to our conceptualization, when learning from experience, the person automatically makes an implicit prediction regarding the future outcome (Friston, 2005; Hohwy, 2013). This prediction could be a random sample from the existing representation. Realized outcomes that produce a prediction-error will be more accurately stored (Kamin, 1969; Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972). Note that the experience format, where outcomes are presented sequentially, naturally allows for prediction error whereas the description format does not. According to our theory, this is the main source of format-dependent differences in judgment and decisions.

We speculate that recollection of the past outcome distribution is produced by a process of randomly sampling from the (imperfectly represented) outcome distribution (Juslin et al., 2007). Similarly, we speculate that generating an expected future outcome distribution is produced by a process of randomly sampling from a *smoothed* version of the (imperfectly represented) outcome distribution. This smoothing process conservatively redistributes the outcome distribution to account for idiosyncrasies in the observed data by making it less "lumpy". For example, a worker may have produced 18 widgets twice last week, which is well above his average of 9. The smoothing process recognizes that these outliers could have very easily been 17 and 18 widgets and so this outcome in the future should not be very surprising. Naturally, this smoothing process is informed by any knowledge of the underlying mechanism producing the outcomes. Support for this idea can be seen in Figure 3 by noting the differences between recalled and predicted outcome distributions. For example, in Panel B, regardless of format, participants indicated that the probability of the extreme 18 outcome was less likely in the future than it was in the past. Importantly, in Experiment 2, choices were better predicted by future outcome predictions rather than the recollection of the presented outcome distribution.

Finally, we propose that choice is determined by the selection of the option with the highest expected value, which is derived directly from the *smoothed* version of the (imperfectly represented) outcome distribution for each option. Support comes from the finding that when this choice rule made a prediction it was very often correct: 87% in Experiment 1, 87% in Experiment 2, and 70% in Experiment 3. This connection between choice and predicted outcomes supports our assumption – one that is consistent with several other recent models – that both estimates and choices emerge from the same underlying representations (Kiani & Shadlen, 2009; Merkle & Van Zandt, 2006; Moran et al., 2015; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2013; Van den Berg et al., 2016). Moreover, choices were equally well predicted in all of our experiments irrespective of the initial learning format. This suggests that format-dependent differences in choice stem from the representation of the information rather than how that representation is applied.

Under this account, the description-experience choice gap emerges because those learning from description have a relatively poorer representation of the future outcome distribution. In particular, they have a much stronger belief in rare, extreme outcomes occurring in the future, at least in contexts with normal or skewed outcome distributions. For example, in Experiment 2, those learning from description predicted extreme events 22% of the time whereas those learning from experience predicted extreme events 14% of the time. A higher anticipation of a rare event changes the expected value calculation.

This sketched account could be investigated further by examining elements of the decision and memory processes thought to be involved. In particular, encouragement to consider more or varied samples (Hayes, Hawkins, & Newell, 2016; Koriat, Lichtenstein, & Fischhoff, 1980; Walters, Fernbach, Fox, & Sloman, 2016), and individual differences in working memory capacity (Dougherty & Hunter, 2003a, 2003b; Kareev, 1995; Rakow, Demes, & Newell, 2008). Future research is also encouraged to test contexts that are different from the hiring manager scenario used in these experiments.

Practical Implications

The primary implication of this research is that people are better able to learn about an outcome distribution when information about it is presented in an experience format compared to a description format. Better calibration allows those who have learned from experience to make relatively more predictions about the future that "hit" the mark. In situations where the underlying outcome distribution is normal or skewed, which we would argue is most situations, then those learning from experience are also less likely to overestimate the possibility of extreme outcomes.

An obvious lesson for those making predictions is to try, wherever possible, to learn from experience-based information. For example, there have been several demonstrations of the benefits of sample simulators (Hawkins et al., 2015; Hogarth & Soyer, 2011; Kaufmann et al., 2013). However, care must be taken because some contexts may be less amenable to applying experience than others, such as when the elements of the initial experience represent only a subset of the elements to be predicted (Hogarth & Soyer, 2016). Another potentially fruitful approach is to encourage joint decisions between an individual who has learned from description and another person who has learned from experience. Initial

evidence suggests that such a "wisdom of the crowd" approach may be helpful in attenuating individual biases (Lejarraga & Müller-Trede, 2016).

Previous research has highlighted the distinction between choosing the best overall option and exceeding a stretch performance target (Kutzner, Read, Stewart, & Brown, 2016). A stretch performance target refers to obtaining an outcome well above average. When seeking to meet a stretch target the variance of performance, in addition to the mean performance, becomes important. For example, to achieve a stretch target the best option may be one with high variance, even if on most occasions that option produces fewer returns than a low variance option. An implication of our work is that estimated variance is moderated by how information is acquired. Therefore, managers pursuing a stretch target may be more likely to prefer candidates learned about from description than comparable candidates learned about by experience because of differences in predicted future performance variance.

Managers are often the recipients of advice that must be integrated to make a decision. Previous research suggests that advisers who learn from description (vs. experience) provide their advice relatively more confidently, and this advice is often more preferred by decision-makers (Benjamin & Budescu, 2015). In this study, decision-makers were not told the format by which advisers learned their information. An implication of the current research is that managers' decisions could be improved if they are trained to ask their advisers how advice-related information was learned and discount the advice from advisors who have learned from description.

Conclusions

Our observations suggest that learning from experience – that is, from sequentially observed outcomes – leads to relatively better encoding and inferences about the properties of the outcome distributions underlying alternative options. This superiority produces relatively better calibrated predictions compared to when learning from description – that is, from a stated summary of the outcome distributions. Most people's behavior is consistent with a choice rule that simply selects the option with the highest predicted expected value. Given that learning from experience tends to produce a lower

expectation of extreme events, those learning from experience sometimes prefer options with better outcomes most of the time.

References

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement Under Uncertainty: Heuristics and Biases* (pp. 294-305). Cambridge.

Aluise, S. J. (2012). Frank Blake has brought Home Depot home: Resurrecting the chain's focus on employee expertise and great customer service is paying off. Retrieved from http://investorplace.com/2012/01/frank-blake-has-brought-home-depot-home/#.V2zRDj8sL4c

Ashby, N. J., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1153.

Barron, G., Leider, S., & Stack, J. (2008). The effect of safe experience on a warnings' impact: Sex, drugs, and rock-n-roll. *Organizational Behavior and Human Decision Processes*, *106*, 125-142. doi:10.1016/j.obhdp.2007.11.002

Barron, G., & Yechiam, E. (2009). The coexistence of overestimation and underweighting of rare events and the contingent recency effect. *Judgment and Decision Making*, *4*, 447-460.

Benjamin, D., & Budescu, D. V. (2015). Advice from experience: Communicating incomplete information incompletely. *Journal of Behavioral Decision Making*, *28*(1), 36-49.

Camilleri, A. R. (2017). The presentation format of review score information influences consumer preferences through the attribution of outlier reviews. *Journal of Interactive marketing*, *38*, 1-14.

Camilleri, A. R., & Newell, B. R. (2009). The role of representation in experience-based choice. *Judgment and Decision Making*, *4*, 518-529.

Camilleri, A. R., & Newell, B. R. (2011). Description- and experience-based choice: Does equivalent information equal equivalent choice? *Acta Psychologica, 136*, 276–284.

Camilleri, A. R., & Newell, B. R. (2013). Mind the gap? Description, experience, and the continuum of uncertainty in risky choice. In V. S. C. Pammi & N. Srinivasan (Eds.), *Progress in Brain Research* (Vol. 202, pp. 55-72). Amsterdam, The Netherlands: Elsevier.

CNBC. (2009). Portfolio's worst American CEOs of all time. Retrieved from http://www.cnbc.com/2009/04/30/Portfolios-Worst-American-CEOs-of-All-Time.html?slide=5

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New York: Wiley.

Den Ouden, H. E. M., Kok, P., & De Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology, 3*, 548.

Dougherty, M. R. P., & Hunter, J. E. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica, 113*(3), 263-282.

Dougherty, M. R. P., & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & cognition, 31*(6), 968-982.

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From Anomalies to Forecasts: Toward a Descriptive Model of Decisions Under Risk, Under Ambiguity, and From Experience. *Psychological Review, 124*(4), 369-409.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., . . . Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making, 23*(1), 15-47.

Erev, I., Glozman, I., & Hertwig, R. (2008). What impacts the impact of rare events. *Journal of Risk and Uncertainty, 36*, 153-177. doi:10.1007/s11166-008-9035-z

Fox, C. R., & Hadar, L. (2006). "Decisions from experience" = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making, 1*, 159-161.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences, 360*(1456), 815-836.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review, 102*(4), 684.

Glöckner, A., Hilbig, B. E., Henninger, F., & Fiedler, S. (2016). The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *Journal of Experimental Psychology: General, 145*(4), 486.

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making, 9*(1), 1.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, 523-551.

Gottlieb, D. A., Weiss, T., & Chapman, G. B. (2007). The format in which uncertainty information is presented affects decision biases. *Psychological Science, 18*, 240-246.

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making, 5*(7), 467.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*(3), 356.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: the case of frequency of occurrence. *American Psychologist, 39*(12), 1372.

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making, 21*(5), 493-518.

Hawkins, G. E., Camilleri, A. R., Heathcote, A., Newell, B. R., & Brown, S. D. (2014). *Modeling probability estimates and choice in decisions from experience.* Paper presented at the Proceedings of the 36th Annual Conference of the Cognitive Science Society, Austin, TX: Cognitive Science Society.

Hawkins, G. E., Hayes, B. K., Donkin, C., Pasqualino, M., & Newell, B. R. (2015). A Bayesian latent-mixture model analysis shows that informative samples reduce base-rate neglect. *Decision, 2*(4), 306.

Hayes, B. K., Hawkins, G. E., & Newell, B. R. (2016). Consider the alternative: The effects of causal knowledge on representing and using alternative hypotheses in judgments under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 723-739.

Hertwig, R. (2012). The psychology and rationality of decisions from experience. *Synthese, 187*(1), 269-292. doi:10.1007/s11229-011-0024-4

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534-539.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences, 13*(12), 517-523.

Hertwig, R., Hogarth, R. M., & Lejarraga, T. (2018). Experience and description: Exploring two paths to knowledge. *Current Directions in Psychological Science, 27*(2), 123-128.

Hertwig, R., Pachur, T., & Kurzenhauser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 621–642.

Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Frontiers in Psychology, 6*.

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science, 24*(5), 379-385.

Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: kind experience versus nontransparent description. *Journal of Experimental Psychology: General, 140*(3), 434.

Hogarth, R. M., & Soyer, E. (2016). Kind and Wicked Experience in Marketing Management. *Journal of Marketing Behavior, 2*(2–3), 81-99.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Jessup, R. K., Bishara, A. J., & Busemeyer, J. R. (2008). Feedback produces divergence from prospect theory in descriptive choice. *Psychological Science, 19*(10), 1015-1022.

Juslin, P., Winman, A., & Hansson, P. (2007). The naive intuitive statistician: a naive sampling model of intuitive confidence intervals. *Psychological Review, 114*(3), 678.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. *Punishment and aversive behavior*.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition, 56*(3), 263-269.

Kaufmann, C., Weber, M., & Haisley, E. (2013). The role of experience sampling and graphical displays on one's investment risk appetite. *Management Science, 59*(2), 323-340.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*(5928), 759-764.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107.

Kutzner, F. L., Read, D., Stewart, N., & Brown, G. (2016). Choosing the Devil You Don't Know: Evidence for Limited Sensitivity to Sample Size–Based Uncertainty When It Offers an Advantage. *Management Science, 63*(5), 1519-1528.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making, 24*.

Lejarraga, T., & Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. *Organizational Behavior and Human Decision Processes, 116*(2), 286-295.

Lejarraga, T., & Müller-Trede, J. (2016). When Experience Meets Description: How Dyads Integrate Experiential and Descriptive Information in Risky Decisions. *Management Science*.

Li, S. Y. W., Rakow, T. R., & Newell, B. R. (2009). Personal experience in doctor and patient decision making: from psychology to medicine. *Journal of Evaluation in Clinical Practice, 15*, 993-995.

Lichtenstein, S., Slovic, P., Fischhoff, B., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 551–578.

Lin, D., Donkin, C., & Newell, B. R. (2015). *The Exemplar Confusion Model: An Account of Biased Probability Estimates in Decisions from Description.* Paper presented at the Proceedings of the 37th Annual Conference of the Cognitive Science Society.

Lublin, J. S., Murray, M., & Brooks, R. (2000). Home Depot names GE's Nardelli as new CEO in a surprise move. Retrieved from http://www.wsj.com/articles/SB976051062408860254

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276.

Merkle, E. C., & Van Zandt, T. (2006). An application of the poisson race model to confidence calibration. *Journal of Experimental Psychology: General*, *135*(3), 391.

Moore, D. A., Carter, A. B., & Yang, H. H. (2015). Wide of the mark: Evidence on the underlying causes of overprecision in judgment. *Organizational Behavior and Human Decision Processes*, *131*, 110-120.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502-517.

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. *The Wiley Blackwell handbook of judgment and decision making*, 182-209.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: novel data and a computational account. *Cognitive psychology*, *78*, 99-147.

Newell, B. R., & Rakow, T. (2007). The role of experience in decisions from description. *Psychonomic Bulletin & Review*, *14*, 1133-1139.

Newell, B. R., Rakow, T., Yechiam, E., & Sambur, M. (2016). Rare disaster information can increase risk-taking. *Nature Climate Change, 6*(2), 158-161.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*(1), 53.

Odean, T. (1998). Do investors trade too much? *American Economic Review, 89*(5), 1279-1298.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*(6), 532.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*(1), 29.

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

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes, 106*, 168-179.

Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making, 23*, 1-14.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: recognition memory and motion discrimination. *Psychological Review, 120*(3), 697.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy

(Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Ries, A. (2007). Out of focus: The rise and fall of Robert Nardelli. Retrieved from http://adage.com/article/al-ries/focus-rise-fall-robert-nardelli/116042/

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: Remembering, imagining, and the brain. *Neuron, 76*(4), 677-694.

Silver, N. (2012). *The signal and the noise: why so many predictions fail--but some don't*: Penguin.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 299.

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes, 83*(2), 282-309.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty - Heuristics and biases. *Science, 185*(4157), 1124-1131. doi:DOI 10.1126/science.185.4157.1124

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science, 20*(4), 473-479.

Van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *Elife, 5*, e12192.

Walters, D. J., Fernbach, P. M., Fox, C. R., & Sloman, S. A. (2016). Known Unknowns: A Critical Determinant of Confidence and Calibration. *Management Science*.

Weber, E. U. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change, 77*, 103-120.

Weiss-Cohen, L., Konstantinidis, E., Speekenbrink, M., & Harvey, N. (2016). Incorporating conflicting descriptions into decisions from experience. *Organizational Behavior and Human Decision Processes, 135*, 55-69.

Wulff, D. U., Canseco, M. M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin, 144*(2), 140-176.

Wulff, D. U., Hills, T. T., & Hertwig, R. (2014). Online product reviews and the description–experience gap. *Journal of Behavioral Decision Making, 28*(3), 214-223.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93*(1), 1-13.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General, 124*(4), 424.

Yechiam, E., Barron, G., & Erev, I. (2005). The role of personal experience in contributing to different patterns of response to rare terrorist attacks. *Journal of Conflict Resolution, 49*, 430-439.

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc. Frequency Processing and Cognition* (pp. 21-36). Oxford: Oxford University Press.

Figures

Figure 1

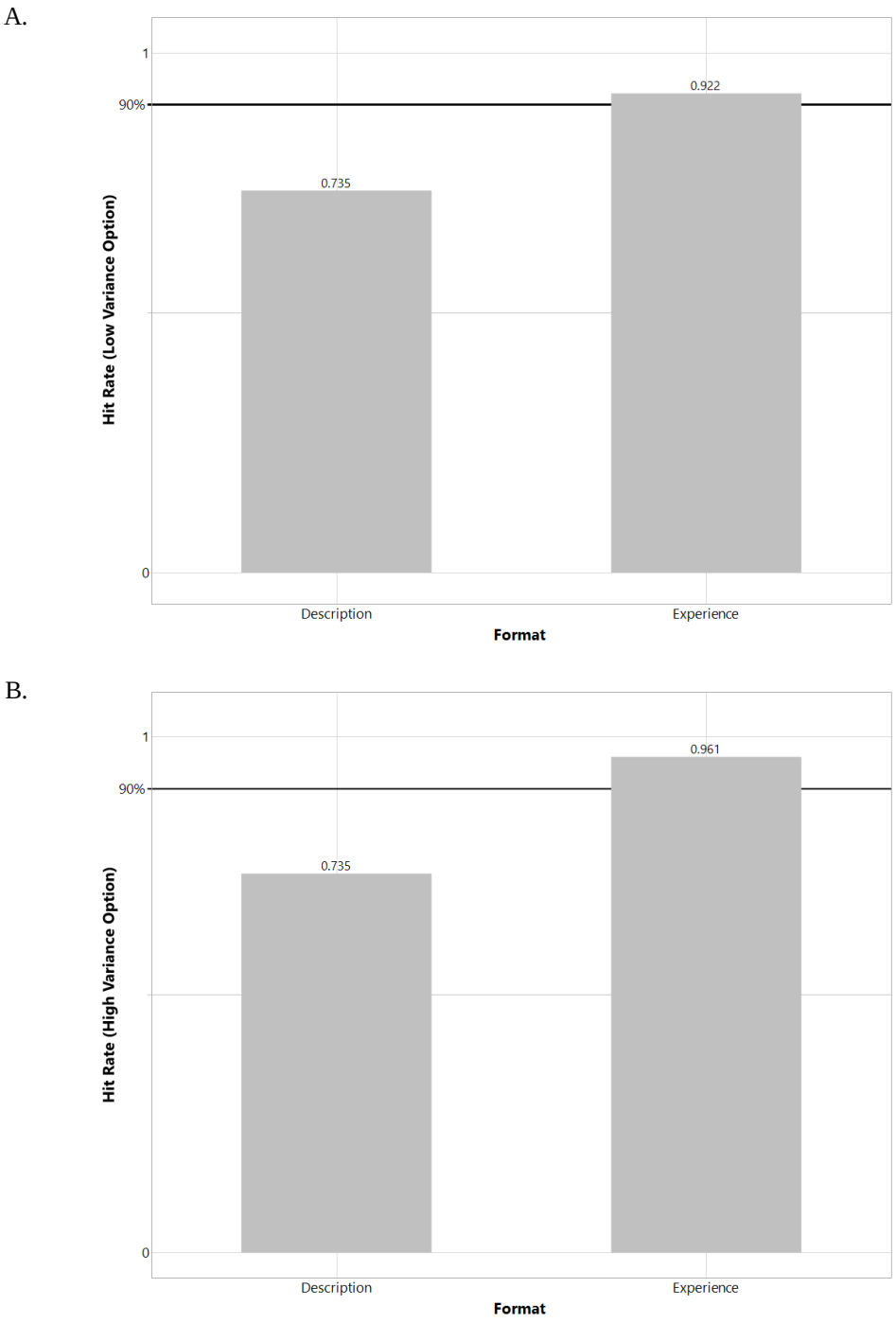
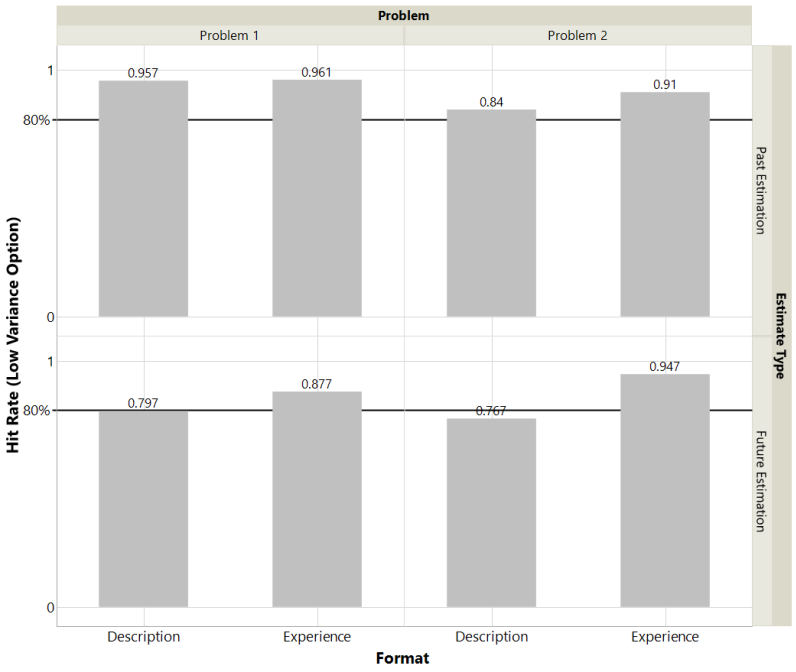


Figure 1. Average proportion of 90% confidence intervals that contain the true mean as a function of format in Experiment 1. A: Low variance option. B: High variance option. The outcome distribution for the low variance option was 8(.5), 9(.5). The outcome disitrbution for the high variance option was 1(.1), 9(.6), 10(.3). The thick horizontal line represents a 90% hit rate.

Figure 2

A.



B.

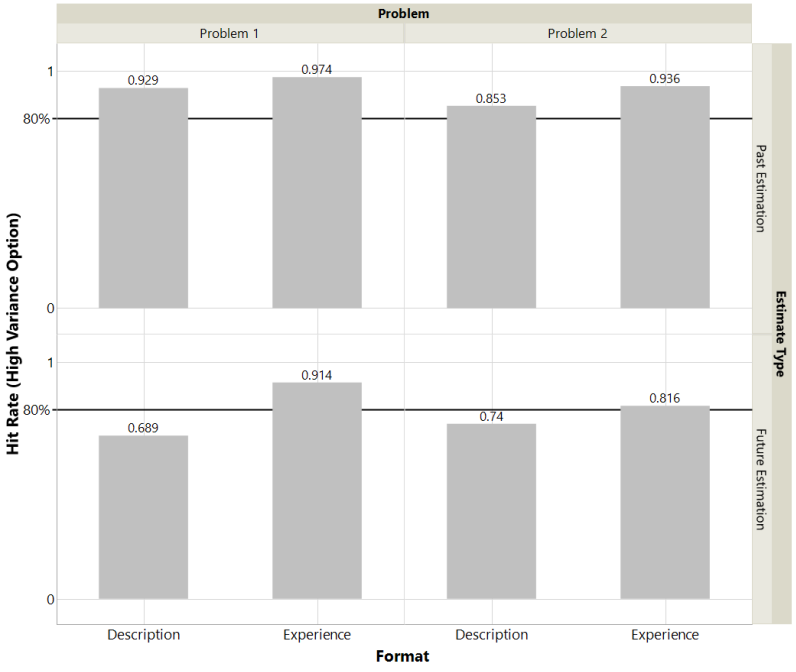
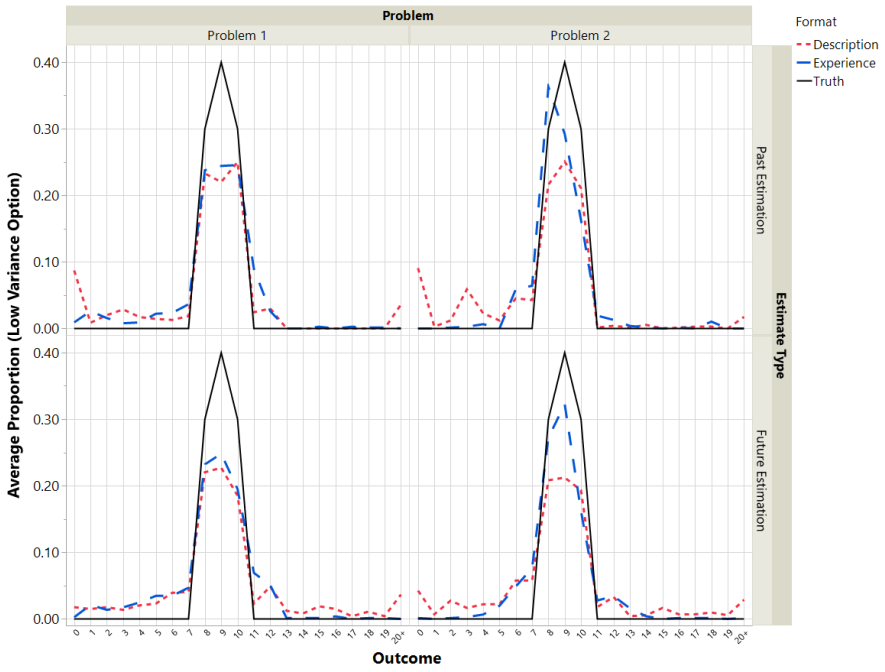


Figure 2. Average proportion of 80% confidence intervals that contain the true mean as a function of format, problem, and estimate type in Experiment 2. A: Low variance option. B: High variance option. The outcome distribution for the low variance option was 8(.3), 9(.4), 10(.3) for all problems. The outcome disitrbution for the high variance option was 1(.2), 10(.2), 11(.4), 12(.2) for Problem 1, and 6(.4), 7(.2), 8(.2), 18(.2) for Problem 2. The dark horizontal line represents 80% hit rate.

Figure 3

A.



B.

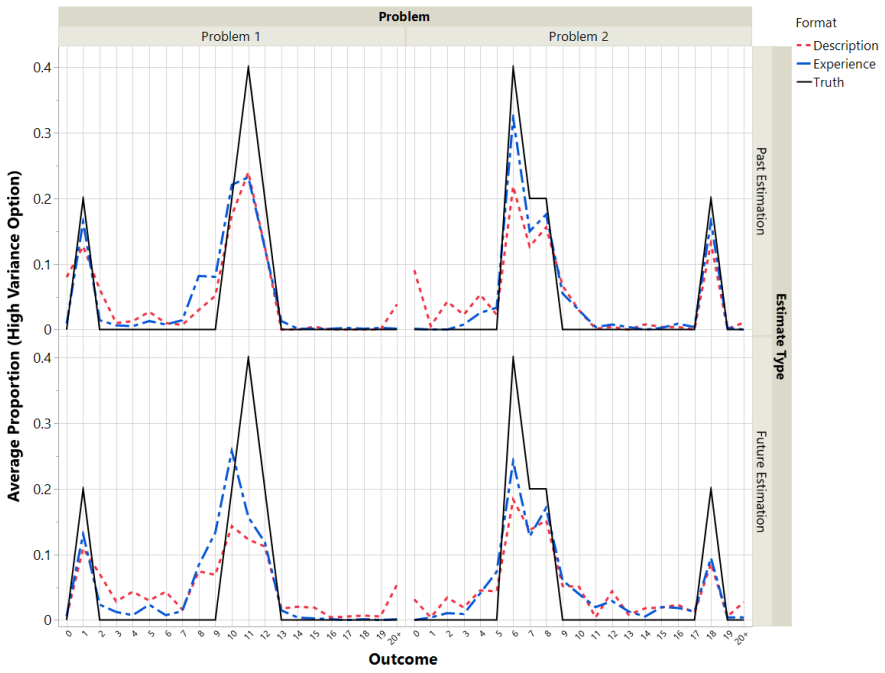


Figure 3. Average estimated likelihood of each outcome as a function of format, problem, and estimate type in Experiment 2. A: Low variance option. B: High variance option. The outcome distribution for the low variance option was 8(.3), 9(.4), 10(.3) for all problems. The outcome disitrbution for the high variance option was 1(.2), 10(.2), 11(.4), 12(.2) for Problem 1, and 6(.4), 7(.2), 8(.2), 18(.2) for Problem 2. Note: values above 20 were rounded down to 20 in this figure.

Figure 4

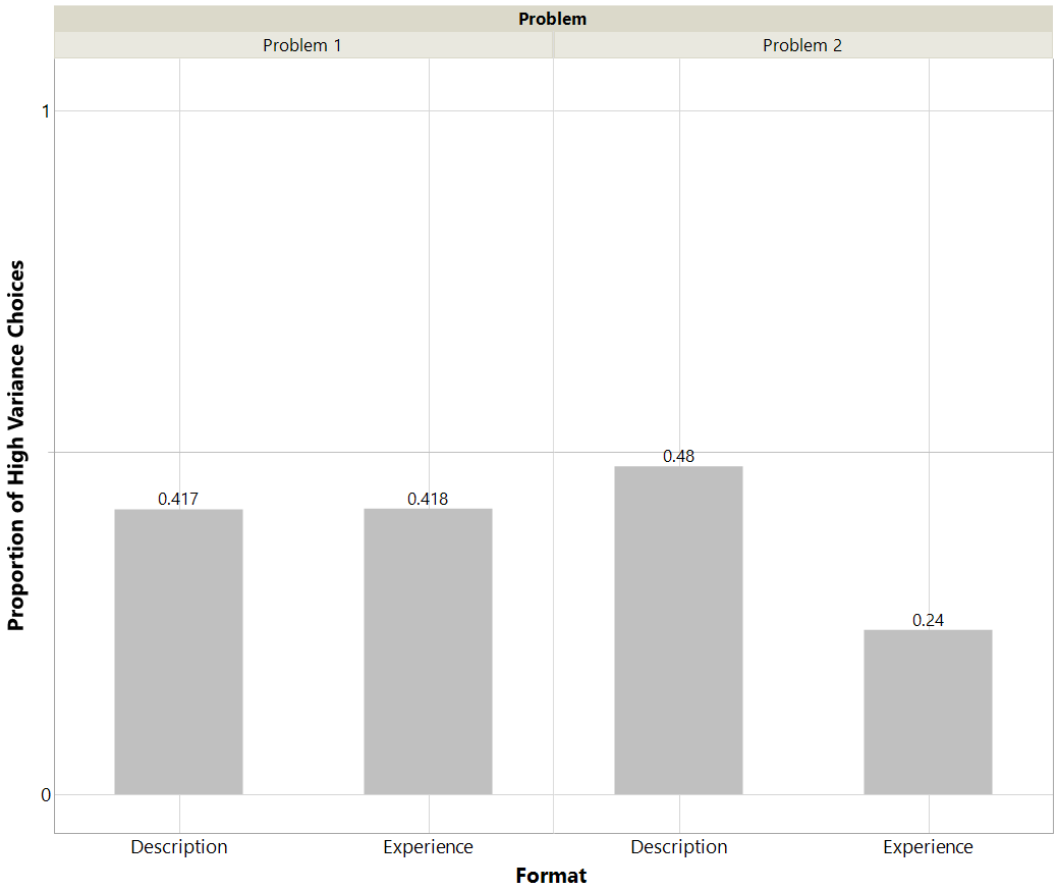


Figure 4. Proportion of high variance choices as a function of format and problem in Experiment 2. The outcome distribution for the low variance option was 8(.3), 9(.4), 10(.3) for all problems. The outcome disitrbution for the high variance option was 1(.2), 10(.2), 11(.4), 12(.2) for Problem 1, and 6(.4), 7(.2), 8(.2), 18(.2) for Problem 2.

Figure 5

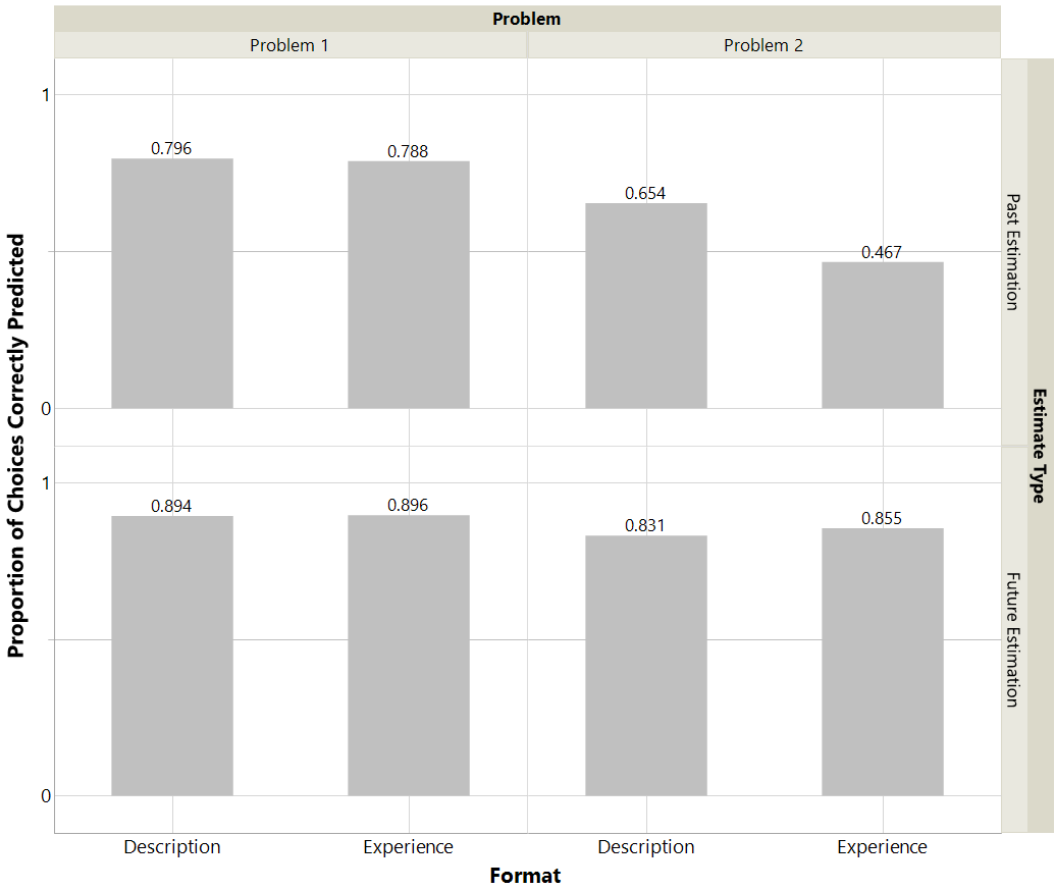
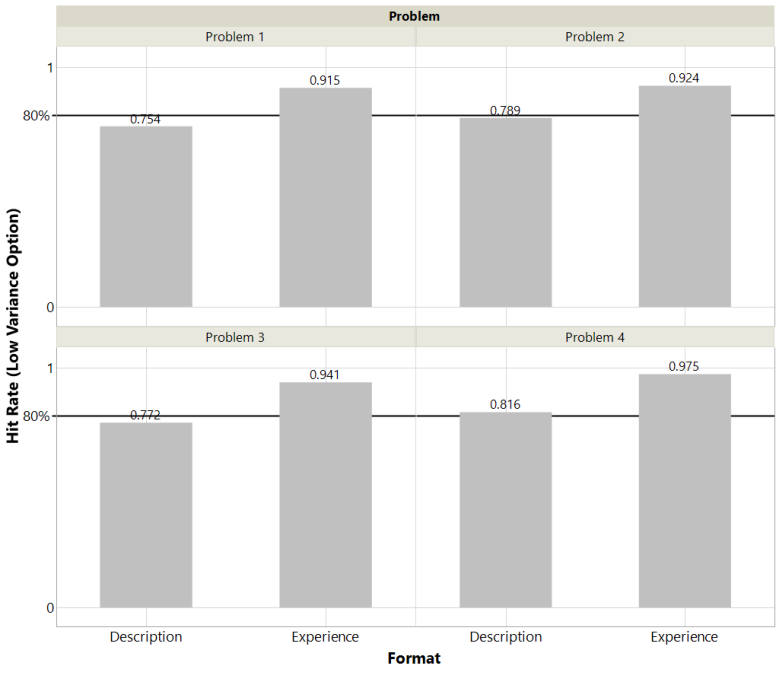


Figure 5. Proportion of choices correctly predicted by the direction of the difference between the implied mean of the distribution of outcomes for the low and high variance options as a function of format, problem, and estimate type in Experiment 2.

Figure 6

A.



B.

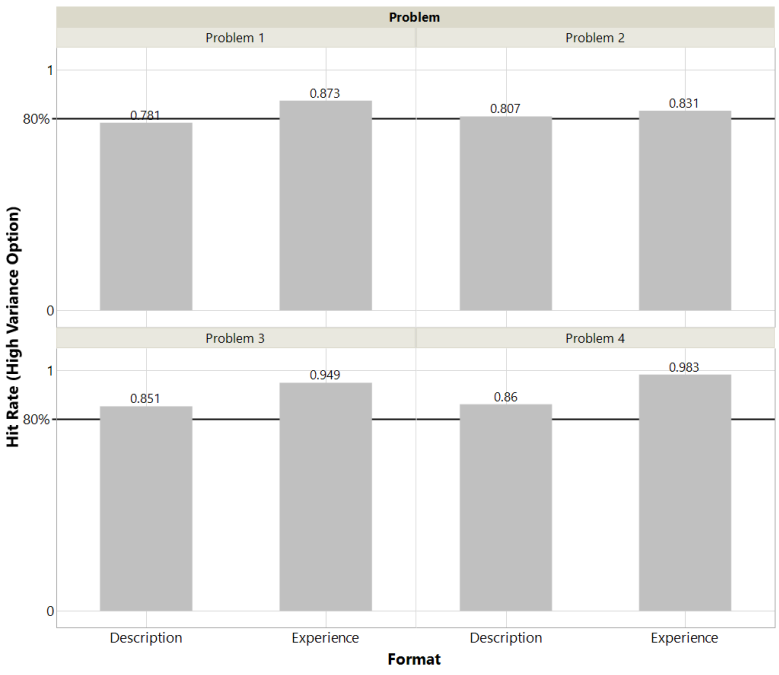
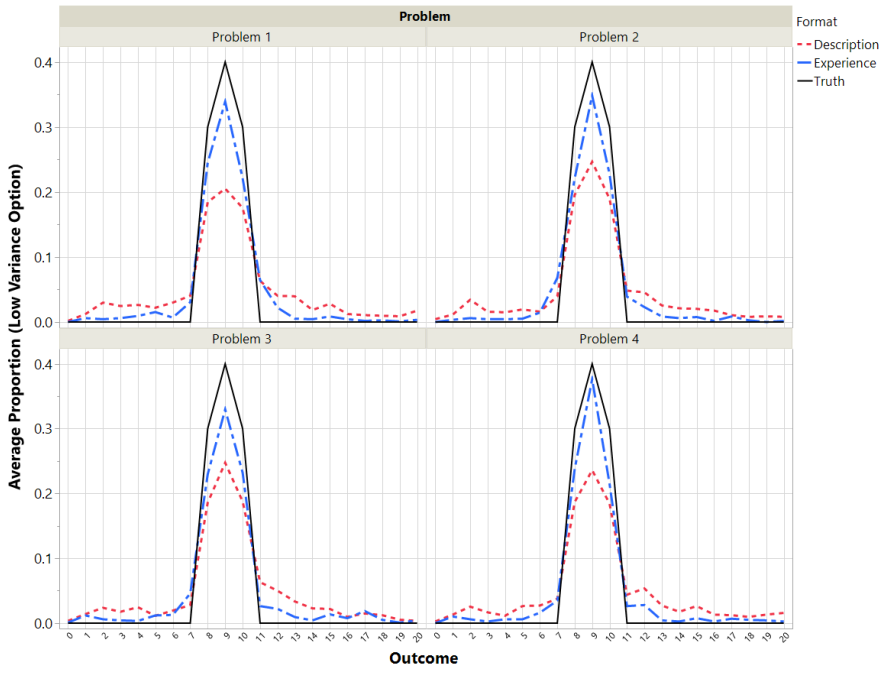


Figure 6. Average proportion of 80% confidence intervals that contain the true mean as a function of format and problem in Experiment 3. A: Low variance option. B: High variance option. The outcome distribution for the low variance option was 8(.3), 9(.4), 10(.3) for all problems. The outcome disitrbution for the high variance option was 1(.2), 11(.8), 12(.2) for Problem 1, 7(.8), 17(.2) for Problem 2, 1(.4), 9(.2), 17(.4) for Problem 3, and 1(.2), 9(.6), 17(.2) for Problem 4. The dark horizontal line represents 80% hit rate.

Figure 7

A.



B.

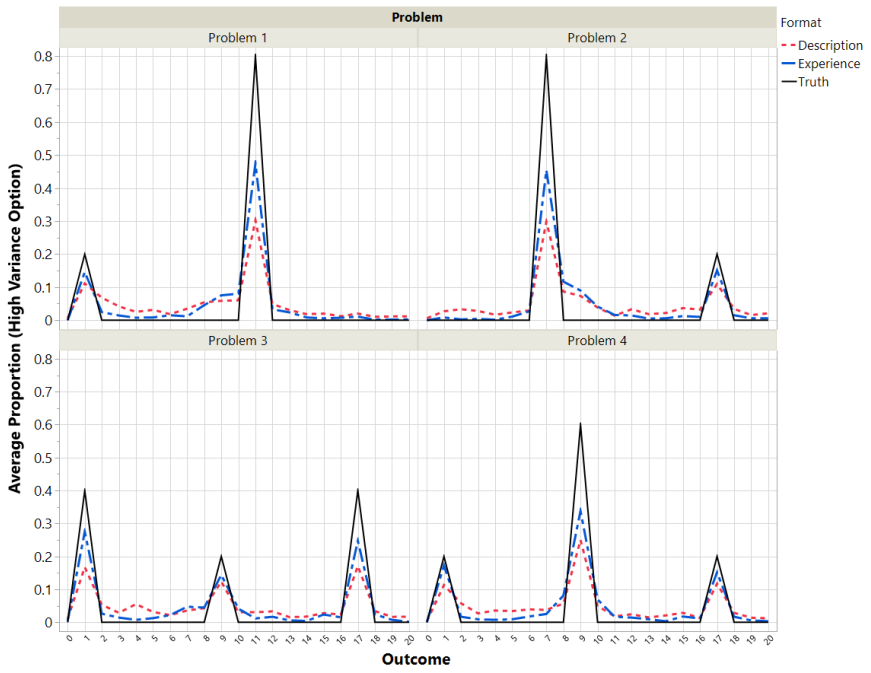


Figure 7. Average estimated likelihood of each outcome as a function of format and problem in Experiment 3. A: Low variance option. B: High variance option. The outcome distribution for the low variance option was 8(.3), 9(.4), 10(.3) for all problems. The outcome disitrbution for the high variance option was 1(.2), 11(.8), 12(.2) for Problem 1, 7(.8), 17(.2) for Problem 2, 1(.4), 9(.2), 17(.4) for Problem 3, and 1(.2), 9(.6), 17(.2) for Problem 4.

Figure 8

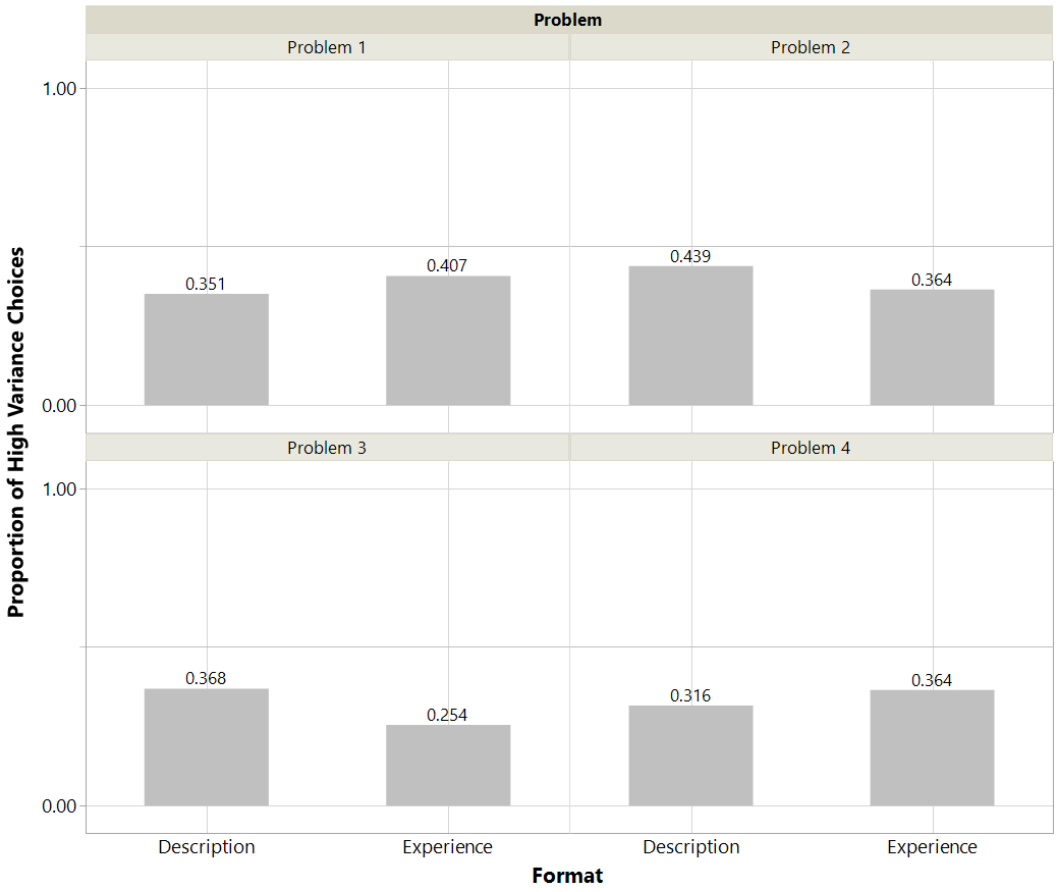


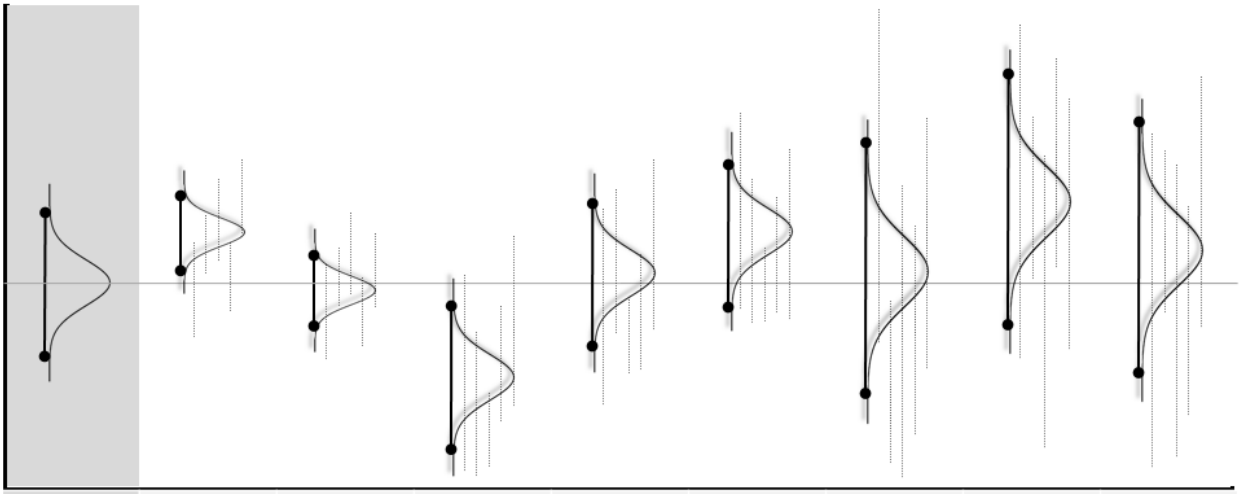
Figure 8. Proportion of high variance choices as a function of format and problem in Experiment 3. The outcome distribution for the low variance option was 8(.3), 9(.4), 10(.3) for all problems. The outcome disitrbution for the high variance option was 1(.2), 11(.8), 12(.2) for Problem 1, 7(.8), 17(.2) for Problem 2, 1(.4), 9(.2), 17(.4) for Problem 3, and 1(.2), 9(.6), 17(.2) for Problem 4.

Appendices

Appendix A

Figure A1 displays the relation between hit rate, interval width, and calibration for eight groups in a hypothetical prediction scenario. Each group comprises 5 participants. The thin dotted lines represent each individual's 80% confidence interval. The solid vertical lines with filled circle caps represent the average 80% confidence intervals for each group. The normal distributions represent the average estimated distribution from each group. The leftmost distribution and interval located in the shaded area indicate the true outcome distribution. The thin horizontal line indicates the mean expected outcome. Hit rate is defined as "precise" (i.e., 80% of intervals contain the mean outcome), "over-precise" (i.e., less than 80% of intervals contain the mean outcome), or "under-precise" (i.e., more than 80% of intervals contain the mean outcome). Interval width is defined as "precise" (i.e., the interval is the same size as the one based on the true outcome distribution), "over-precise" (i.e., the interval is smaller than the one based on the true outcome distribution), or under-precise (i.e., the interval is larger than the one based on the true outcome distribution). Calibration is defined using qualitative labels of "low", "moderate", or "high".

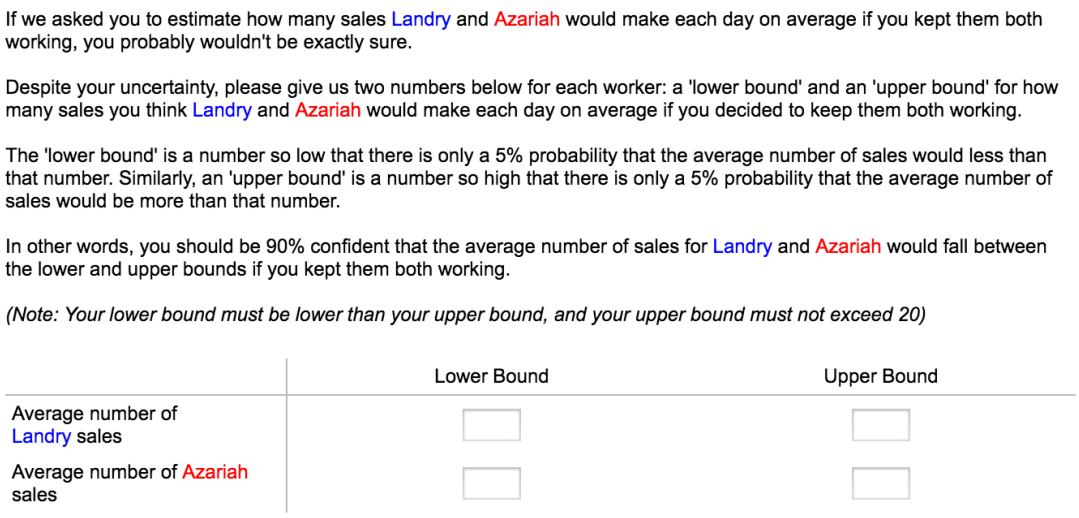
The different groups show possible scenarios in which hit rate and interval width conflict, most notably for Group 2 and 7. For example, those in in Group 2 are considered: (1) under-precise in terms of hit rate because this group's set of 80% confidence intervals includes the mean expected outcome more than 80% of the time (i.e., 5 out of 5 times); (2) over-precise in terms of interval width because the group's average 80% confidence interval width is smaller than the 80% confidence interval derived from the true outcome distribution; (3) good calibration because the group's average predicted probability of each outcome is close to the true probability of each outcome. In contrast, those in in Group 7 are considered: (1) over-precise in terms of hit rate because this group's set of 80% confidence intervals includes the mean expected outcome less than 80% of the time (i.e., 3 out of 5 times); (2) under-precise in terms of interval width because the group's average 80% confidence interval width is larger than the 80% confidence interval derived from the true outcome distribution; (3) poor calibration because the group's average predicted probability of each outcome is far from the true probability of each outcome.



Group	1	2	3	4	5	6	7	8
Hit Rate	Over-precise	Under-precise	Over-precise	Precise	Under-precise	Precise	Over-precise	Under-precise
Interval Width	Over-precise	Over-precise	Precise	Precise	Precise	Under-precise	Under-precise	Under-precise
Calibration	Moderate	Good	Poor	Good	Moderate	Moderate	Poor	Moderate

Figure A1. The relation between hit rate, interval width, and calibration for eight groups in a hypothetical prediction scenario.

Appendix B



If we asked you to estimate how many sales Landry and Azariah would make each day on average if you kept them both working, you probably wouldn't be exactly sure.

Despite your uncertainty, please give us two numbers below for each worker: a 'lower bound' and an 'upper bound' for how many sales you think Landry and Azariah would make each day on average if you decided to keep them both working.

The 'lower bound' is a number so low that there is only a 5% probability that the average number of sales would less than that number. Similarly, an 'upper bound' is a number so high that there is only a 5% probability that the average number of sales would be more than that number.

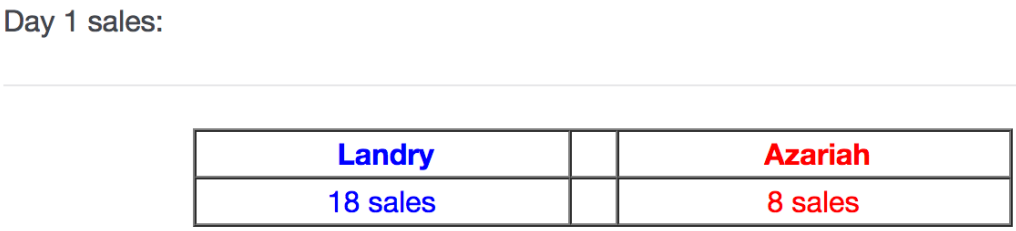
In other words, you should be 90% confident that the average number of sales for Landry and Azariah would fall between the lower and upper bounds if you kept them both working.

(Note: Your lower bound must be lower than your upper bound, and your upper bound must not exceed 20)

	Lower Bound	Upper Bound
Average number of Landry sales		
Average number of Azariah sales		

Screenshot of the question asked of participants to determine their confidence interval around average future sales for each worker in Experiment 1. The lower bound and upper bound were required to be integers between the values of 0 and 20. The lower bound was required to be smaller than the upper bound.

Appendix C



Screenshot of the information provided to participants learning from experience-based information. There were nine other similar screens indicating the performance on each of ten days in both Experiment 1 and 2.

Appendix D

Landry	Azariah
Over the last 10 days, on 50% of days made 8 sales and on the remaining 50% of days made 9 sales.	Over the last 10 days, on 10% of days made 1 sale, on 60% of days made 9 sales, and on the remaining 30% of days made 10 sales.

Screenshot of the information provided to participants learning from description-based information in Experiment 2.

Appendix E

For those in the description-followed-by-experience (henceforth, DE) group, worker performance was first summarily described (as for those in the description group) and then presented as individual outcomes (as for those in the experience group). For those in the experience-followed-by-description (henceforth, ED) group, worker performance was first presented as individual outcomes and then summarily described. Participants in both of these groups were made aware that the description and experience formats communicated the same information.

The mean (and standard deviation of) the lower and upper bounds of the 90% confidence intervals and choice for each group in Experiment 1 is presented in Table A1. For the low variance option, there was no significant difference in *hit rate* between groups, $\chi^2(N = 202) = 6.52$, $p = .09$. However, planned follow-up contrasts revealed that *hit rate* was higher for those in in the experience group than those in the description group, $\chi^2(N = 100) = 6.42$, $p = .01$, but not the DE group ($M = .84$), $\chi^2(N = 102) = 1.55$, $p = .21$, nor the ED group ($M = .84$), $\chi^2(N = 102) = 1.54$, $p = .21$. For the high variance option, there was a significant difference in *hit rate* between groups, $\chi^2(N = 202) = 12.64$, $p < .01$. Planned follow-up contrasts revealed that *hit rate* was higher for those in the experience group than those in the description group, $\chi^2(N = 100) = 10.94$, $p < .001$, but not the DE group ($M = .86$), $\chi^2(N = 102) = 3.18$, $p = .07$, nor the ED group ($M = .92$), $\chi^2(N = 102) = 0.73$, $p = .39$. In addition, choices varied across groups, $\chi2(N = 202) = 10.95$, $p < .05$. Planned follow-up contrasts revealed that the proportion of high variance choices was higher for those in the experience group than those in the description group, $\chi2(N = 100) = 9.22$, $p < .01$, the DE group (.43), $\chi2(N = 102) = 4.99$, $p < .05$, and also the ED group (.39), $\chi2(N = 102) = 6.76$, $p < .001$.

In summary, we observed that those presented with both description and experience information tended to choose similarly to those presented with description-based information alone. This finding is somewhat inconsistent with previous research, which has reported that people tend to rely on experience rather than description-based information when both are available (Barron, Leider, & Stack, 2008; Jessup, Bishara, & Busemeyer, 2008; Lejarraga & Gonzalez, 2011; Weiss-Cohen, Konstantinidis, Speekenbrink,

& Harvey, 2016). One explanation for these somewhat conflicting observations is the amount of experience-based information available. In our studies, the sample was quite limited (i.e., 10 outcomes). In contrast, previous studies have used much larger samples (e.g., 100 in Lejarraga & Gonzalez, 2011). Therefore, it seems likely that experience-based information comes to overwhelm description-based information as more observations are made (Newell & Rakow, 2007).

Table A1

Mean (and Standard Deviation of the) Lower and Upper Bounds of the 90% Confidence Intervals and Choice for Each Group in Experiment 1

	Description	ED	DE	Experience
N	49	51	51	51
	Low Variance Option			
Lower Bound	6.1 (3.2)	6.5 (2.3)	5.9 (2.1)	6.3 (1.8)
Upper Bound	10.1 (3.6)	10.6 (3.3)	10.0 (2.7)	10.0 (2.0)
Hit Rate	.74	.84	.84	.92
	High Variance Option			
Lower Bound	3.5 (3.6)	4.0 (3.1)	3.9 (2.9)	5.4 (2.8)
Upper Bound	10.6 (4.1)	11.5 (3.1)	10.6 (2.8)	10.8 (2.4)
Hit Rate	.74	.92	.86	.96
Proportion Choosing the High Variance Option	.35	.39	.43	.65

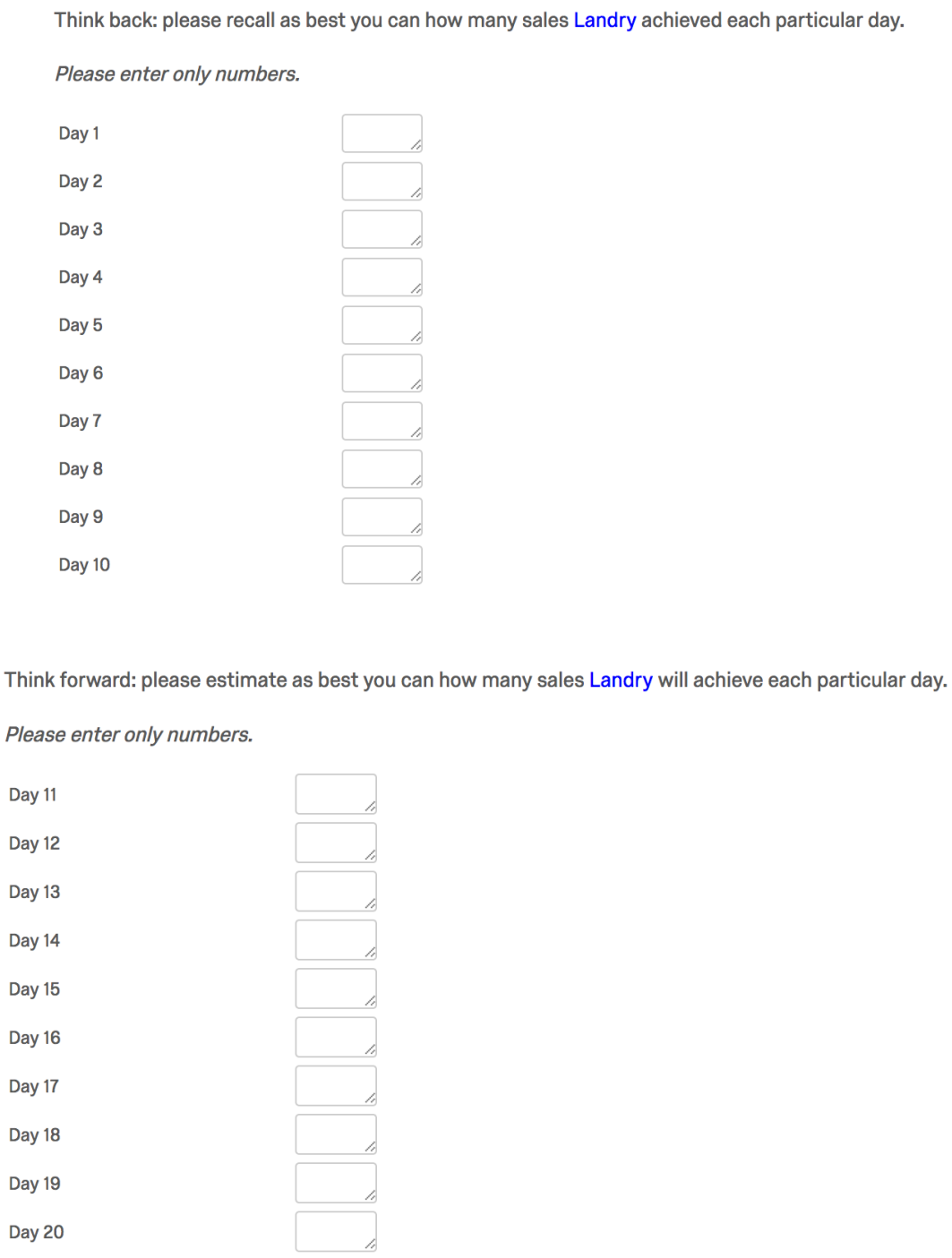
Note: ED refers to the "Experience then description" group; DE refers to the "Description then experience" group.

Appendix F

Landry	Azariah
Over the last 10 days, made 8 sales three (x3) times, 9 sales four (x4) times, and 10 sales three (x3) times.	Over the last 10 days, made 6 sales four (x4) times, 7 sales two (x2) times, 8 sales two (x2) times, and 18 sales two (x2) times.

Screenshot of the information provided to participants learning from description-based information in Experiment 2.

Appendix G



Screenshot of the question asked of participants to determine their estimated outcome distribution in Experiment 2.

Appendix H

Instructions describing incentives associated with choices in Experiment 3:

> *We want you to really engage with this study. Therefore, we are going to put some real cash on the line.*
>
> *For every 20 people who complete this study, we will select 1 of those people to be paid a cash bonus payment equal to the average 10-day future performance of one randomly selected worker you chose to keep. This bonus payment will range between $0 and $20. The winners can collect the cash from the researcher at the conclusion of the study.*

Instructions describing incentives associated with estimates in Experiment 3:

> *In order to ensure you take this task seriously, we are also going to reward you with a potential cash payment based on your judgments. Basically, the more accurate your responses turn out to be, the higher your potential reward. This cash reward will be added to the payment going to the 1-in-20 people who end up being selected at the conclusion of the study.*
>
> *The formula that we will apply to calculate your accuracy is the summed absolute distance between your forecast and the actual distribution for each worker. This formula may appear complicated, but what it means for you is very simple: You get paid the most when you honestly report your best guesses about the expected number of sales for each day and each worker. The range of your payoffs for these judgments is between $0 to $4.*