

Running head: AMOUNT OF INFORMATION

Effects of Amount of Information on Judgment Accuracy and Confidence

Claire Tsai, Joshua Klayman, and Reid Hastie

University of Chicago

Corresponding Author: Claire Tsai
Graduate School of Business
University of Chicago
Chicago IL 60637
Electronic mail: mitsai@gsb.uchicago.edu

KEYWORDS: judgment; confidence; accuracy

Abstract

When a person makes a judgment and assesses confidence, what is the effect of receiving more judgment-relevant information? Following up prior research on belief updating judgments (Oskamp, 1965; Slovic, 1973), we demonstrated that the predictive accuracy of experts did not improve but their confidence climbed steadily as they received more information about the target event. Our findings showed a substantial discrepancy between the judgment and confidence in the judgment. We outline a framework that identifies the key "cues to confidence" in judgments under uncertainty and discuss the implications of the present research for a general theory of confidence in judgment.

Effects of Amount of Information on Judgment Accuracy and Confidence

In many situations, ranging from financial investments and medical dilemmas to poker games, it is our degree of confidence in a decision-relevant judgment that determines the actions we take. But, what psychological mechanisms underlie our intuitive sense of confidence and how trustworthy a guide is confidence when we decide to take a consequential action? In particular, how do we update judgment confidence as additional information is acquired and accumulated over an extended period of time?

The focus of behavioral research on confidence judgments has been on accuracy or, as it is called in this literature, calibration: When someone makes a confidence judgment, how does that confidence relate to the actual probability of occurrence of the judged event or to the accuracy of a point estimate? The most popular research paradigms to study confidence calibration involve the use of questions about the type of facts found in almanacs (What is the longest river in the world? What is the length of the Nile in kilometers?) and predictions of the outcomes of uncertain events (Will the Bears play in the Superbowl this year?). Participants generate their best answer and then assign a subjective probability rating (“I’m 60% sure that AT&T stock will be selling at a higher price this time next year”; Yates, 1990) or a confidence interval (“Give us two numbers such that you are 80% sure that the correct answer [e.g., the invoice price of different sedan-type automobiles] lies somewhere between the two”; Soll & Klayman, 2004).

This research has yielded a trove of important methodological lessons concerning the proper assessment of calibration and some general behavioral conclusions. Many studies conclude that people are overconfident (Einhorn & Hogarth, 1978; McClelland &

Bolger, 1994; Yates, 1990). This has been attributed to motivational factors (Babad, 1987; Kunda, 1990; Langer, 1975; Larrick, 1993) and to cognitive biases such as anchoring and insufficient adjustment, confirmatory reasoning in memory retrieval, and biased interpretation of ambiguous information (Hoch, 1985; Klayman, 1995; Koriat, Lichtenstein, & Fischhoff, 1980; Sherman, Zehner, & Johnson, 1983). However, it has also been demonstrated that in many cases confidence diverges from accuracy simply because people are unable to perfectly assess the predictive validity of judgment-relevant information (Gigerenzer, Hoffrage, & Kleinbolting, 1991; Soll, 1996).

Research also shows that the relationship between confidence and accuracy depends on a number of different variables. Some domains engender overconfidence more readily than others (Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999; Soll, 1996), with some evidence that more difficult domains show greater overconfidence (Ferrell, 1994; Ferrell & McGoey, 1980; Juslin, P., Winman, A., & Olsson, H., 2000; Peterson & Pitz, 1988; Suantak, Bolger, & Ferrell, 1996). Some individuals are more prone to overconfidence, with mixed evidence suggesting males are more so than the females (Barber & Odean, 2001; Lichtenstein & Fischhoff, 1981; Lundeberg, Fox, & Puncochar, 1994; Pallier, 2003; Soll & Klayman, 2004). In general, experts seem to be better calibrated and less overconfident than novices (Koehler, Brenner, & Griffin, 2002). Confidence judgments concerning the self seem more prone to overconfidence than judgments concerning others (Griffin, Dunning, & Ross, 1990; Harvey, Koehler, & Ayton, 1997). The format of the confidence assessment also has a large impact. For example, confidence in choices between two alternatives is fairly accurate whereas

confidence in setting an interval around an estimate often greatly exceeds accuracy (Juslin, Wennerholm, & Olsson, 1999; Klayman et al., 1999).

Although the complexity of calibration results may seem daunting, we believe that they can be understood in terms of judges using multiple cues to infer confidence. We agree with many theorists (Erev, Wallsten, & Budescu, 1994; Gigerenzer et. al., 1991; Juslin & Olsson, 1997; Yates, 1990) that the primary cues to confidence are impressions of the validity of the available information, but we propose that the amount of information (completeness), the consistency of the information (coherence), and the ease of the judgment experience (fluency) are also important in assessments of confidence. Since these latter cues are often not correlated with potential accuracy, reliance on them to assess confidence can produce accuracy-confidence discrepancies. In the present research we focus on two cues that we hypothesize to be important in determining the relationship between confidence and amount of available information; impressions of the completeness and the coherence of judgment-relevant evidence.

Perceived completeness is primarily a function of the amount of information that has been acquired relative to the amount of all the available judgment-relevant evidence. On the other hand, subjective cue validity might also influence the perceived completeness of information. For example, knowledgeable judges might consider certain cues to be more diagnostic than others and their perceived completeness of the information can be enhanced if the strong (i.e., highly valid) cues are available to them and vice versa.

Perceived coherence is another important source of increasing confidence as information accumulates. The evidence on this point is primarily from studies of legal

judgments, where facilitating the creation of a narrative and providing additional, non-diagnostic details increases confidence (Bell & Loftus, 1989; Pennington & Hastie, 1986, 1988). We hypothesize that judges might find it easier to form or evaluate a hypothesis as additional information is acquired even though the newly acquired information does not provide additional explanatory power.

Only a few prior behavioral studies have examined confidence in belief updating tasks. Oskamp (1965) used a clinical psychology case study to demonstrate that judges (including practicing clinical psychologists) showed increasing overconfidence in their judgments as more case information was presented. However, the quantity and statistical properties of the case information were ill-defined, so that properties such as the correlation, diagnosticity, or consistency among the cues was unspecifiable. A methodological advance was provided by an unpublished study of horse racing handicappers. Slovic and Corrigan (1973; report of a study conducted with B. Corrigan; cited in a popular textbook by Russo & Schoemaker, 2002, pp. 123-124) showed that the predictive accuracy of racing touts did not improve as they received more statistical cues about harness race contestants, although their confidence climbed steadily over time. From a set of 40 cues, participants selected the specific cues they wanted to study in consecutive blocks of 5, 5, 15, and 15 cues each. In both these studies, the confidence-accuracy discrepancy seems to occur because participants did not realize that their accuracy hit a ceiling after the first set of information. Peterson and Pitz (1986, 1988) also manipulated amounts of information presented, using one, two, or three valid, non-redundant statistical cues. Their participants judged baseball teams' performance (win/loss, number of wins in a season). In general, confidence and accuracy both

increased with the presentation of more information, although accuracy did not increase as fast as confidence.

These previous findings also raise the question of why accuracy may begin to level off at an early point in the sequential judgment process. Slovic and Corrigan (1973) suggests that information overload might impose a ceiling on accuracy, as limited cognitive capacity may interfere with the processing of new information as the amount of information increases. Flat accuracy can occur if judgments become more noisy as they attempt to process larger amounts of information. Another element is irreducible uncertainty, that is, that there is some unpredictability that remains in any real-world judgment no matter how much information is acquired (Hammond, 1996, pp. 13-35). We hypothesize that judges are less sensitive to elements such as these than they are to cues like apparent completeness and coherence of information. Thus, confidence in judgments and decisions may continue to rise long after the actual accuracy has leveled off.

Experiment 1

In this study, we sought to examine judgment accuracy and confidence in a sequential information acquisition task with knowledgeable judges using real-world data. Each participant was asked to predict the winner and point spread of 15 NCAA college football games given statistical information about the two competing teams. These are familiar types of judgments for football fans: It is common for fans to look up statistics from sports web sites before a game and then place bets based on their predicted winners and point spreads. For each game, cues were provided in five blocks of six cues each. After each block, participants made predictions about the game and assessed their confidence in their predictions.

Method

Participants. Participants were 30 college and graduate students at the University of Chicago. On average, participants spent about one hour to complete the experiment in exchange for a fixed payment of \$15. In addition, a reward of \$50 was promised to the participant with the best performance. In order to take part in the study, participants had to pass a test demonstrating they were highly knowledgeable about college football.

Task. We wanted a task that afforded a moderate degree of prediction from a number of measurable cues. Statistical analysis has shown that football is a relatively predictable sport compared with baseball, basketball, and many other sports. We measured task predictability on two criterion values including predictions of winners and estimates of point spreads. Stepwise, logit regression models were constructed using a hold-out sample of games.¹ Then we applied the models to predict the same games using the same blocks of cues presented to participants. Correct selection of the winner improved as the number of variables provided to the models increased. Models were correct 56% of the time with six cues and 71% with 30 cues (see Figure 1).

Similarly, accuracy of point spreads estimates (measured as the difference between the estimated predictions and actual point spreads) also improved as additional cue blocks were included in the stepwise multiple regression models from a mean absolute deviation (MAD) of 14 with 6 cues to a MAD of 12 with 30 cues. Overall, the increase in predictive accuracy of the regression models confirms that college football is relatively predictable, at least in principle, and that the barrier of irreducible uncertainty has not been reached prior to the last block of cues.

Design. Trial was a repeated measure, with six additional cues being provided on each trial. We counterbalanced the order of cue blocks between participants using a 5x5 Latin-Square, and the order of blocks was held constant for each individual participant. Each participant received 15 matches, presented in one of two different orders, with half of the participants receiving the games in each order. The order of games was set by starting with a random ordering, and then adjusting to avoid streaks of either very close games or routs.

Stimuli. We obtained 45 matches by randomly sampling the end-of season conference games from NCAA football seasons in 2000-2002. We selected only from among conference games (i.e., those between teams in the same conference). Comparative statistics within a given conference are more easily interpreted, because team-to-team differences are not confounded with differences between conferences. In order to enhance the representativeness of our sample, several different random samples were drawn, and we selected one of them for which the distribution of point spreads and the proportion of upsets most closely matched that for all college football games played at the end of the season in 2000-02.² The 45 matches were then divided into three subsets of 15 each, with each subset also being representative in these ways. Each subset was presented to ten participants.

The names of the teams were not provided, and were labeled A and B. To minimize the chance that participants might learn to favor one letter over the other, letters were assigned such that Team A and Team B each won about half the games, by the same average margin.

The 30 football performance statistics were selected based on their subjective cue validity. Perceived cue validity of the football statistics was pretested by asking a separate sample of college football fans to rank order 106 football statistics downloaded from the official NCAA football web site. These cues were divided into three categories – highly, moderately, and least valid cues – based on the average rankings. The 30 cues include the ten rated as most valid, the ten least valid, and ten moderately valid cues (those ranked from 49th to 58th). This set of 30 cues was also used in Experiments 2 and 3.

Procedure. At the beginning of the experiment participants were told the purpose of the study was to understand their strategies to predict college football. Subsequently, they were asked to predict the outcome of 16 college football matches and were not aware that the first match was practice and that it did not count for their overall performance. All participants received the same practice match.

Instructions and stimuli were presented by computer. After the presentation of each new block of six cues, participants were asked to pick the winner and estimate the point spread for NCAA football matches based on the available statistics. Subsequently, they indicated the judgment confidence associated with their predictions. The instructions provided for this are shown in Figures 2 and 3.

For each match, 30 football statistics for the paired teams were presented in a box table that resembled the display format of football statistics on ESPN and most popular sports web sites. Participants were given a block of six cues at a time; hence, there were five blocks per match. Information from previous blocks stayed on the screen when a new block of six cues were presented. Each time additional information was presented,

participants were asked to pick the winner and assess their confidence by indicating the chance that their choice was correct, ranging from 50 to 100%. Then they were asked to estimate the point spread and indicate their confidence by giving the upper and lower bound of a 90% confidence interval. They repeated this process until they received all 30 cues. When participants received all 30 cues and finished the tasks for the last trial, participants were given another match to predict. They continued this process until they completed the prediction tasks for all 16 matches.

Results

Except where noted, judgments were analyzed using a MANOVA with Trial (1 – 5) as a within-participants variable. Trial number indicates the amount of information available to the participants (6, 12, 18, 24, and 30 cues for trials 1 to 5, respectively). The dependent measures for prediction of winners were the proportion of correct predictions and the subjective probability of being correct, averaged across the 15 games for each participant. For point spread estimate, the dependent variables were the absolute difference between the predicted and actual point spreads and the width of the 90% confidence intervals.

Picking the winner. As shown in Figure 4, accuracy remained flat as the amount of information increased. A test for linear trend³ did not yield a significant result, $F(1, 28) = 0.78, p = 0.38$. The 66% accuracy achieved on the last trial is significantly above chance, $t(29) = 12.81, p < 0.001$, though less than the 71% hit rate by the statistical model, $t(29) = -3.83, p < 0.001$.

Confidence rose steadily as additional information was presented (see Figure 4). We observed a significant linear increase in overall confidence from trial 1 to trial 5, $F(1,$

28) = 43.39, $p < 0.001$. As predicted, the discrepancy between confidence and accuracy also increased linearly across trials, $F(1, 28) = 14.47, p < 0.001$.

Estimating point spreads. For this task, we measured accuracy by calculating the mean absolute differences (MAD) between the estimated and actual point spreads.⁴ As shown in Figure 5, MAD was flat across trials, $F(1, 28) = 3.23, p < 0.08$.⁵ Confidence in estimating point spreads was measured by the width of the confidence intervals.⁶ The narrower the intervals is, the greater the implied confidence. We found confidence rose markedly as new information was presented (see Figure 5). A trend analysis showed a significant linear trend, $F(1, 28) = 10.48, p = 0.003$, as well as a non-linear trend, $F(1, 28) = 6.53, p = 0.02$.

In order to compare accuracy with confidence, we applied a linear transformation to the confidence interval widths to obtain the mean expected absolute deviation (MEAD; Soll & Klayman, 2004). MEAD is an estimate of the absolute deviation between true and estimated values that would be expected on average were the intervals well calibrated.⁷ Following the method proposed by Soll and Klayman (2004), we measured overconfidence by taking the ratio, $M = \text{MEAD}/\text{MAD}$. If judges are perfectly calibrated, $\text{MEAD} = \text{MAD}$ and, thus, $M = 1$. On the other hand, if judges are overconfident, their confidence intervals and MEAD are too small and $M < 1$. Figure 5 shows that MEAD was constantly less than MAD. A test for linear trend for M was significant, $F(1, 28) = 3.94, p = 0.05$. Judges became more confident, reducing their confidence intervals as they acquired additional information over time, while their accuracy was actually declining slightly. Thus, point spread estimates also show the predicted pattern of increased overconfidence with more information.

Experiment 2

A potential limitation of Experiment 1 is that expressions of confidence are not tied to any real consequences for participants. Moreover, confidence was expressed using numerical scales that, while common in experimental methods and decision analysis, may be unfamiliar to participants. Thus, our next study tested the effects of accuracy and confidence with real money at stake. Similar to Experiment 1, participants were given information about NCAA football teams, in five blocks of six cues each, and were asked to predict the outcome of 30 college football games. After each judgment they were asked to either place a monetary value for betting on their answer or to match such a bet to the odds of winning in a lottery. Payoffs were determined by their judgments.

Method

Participants. Participants were 20 college and graduate students from the University of Chicago and they passed the football knowledge test to take part in this experiment in exchange for a wage of \$5 and had a chance to win a bonus up to \$15. The experimental session lasted about 90 minutes. In addition, a reward of \$50 was promised to the participant with the best performance in this football forecast contest.

Design, Task, Stimuli, and Procedure. The tasks and stimuli used in this experiment were identical to those used in Experiment 1, except for three modifications. First, two subsets of games were selected from the three sets used in Experiment 1. The distribution of upsets, the percentage of Team A being the winner, and the order of the games within each subset were identical to Experiment 1. Unlike in Experiment 1, participants were asked to predict both sets of games, for a total of 30 games each.

Second, participants were asked to predict only the winning teams, not the point spreads. Lastly, the hypothetical confidence rating (subjective probability of correct winner selection) was replaced with two behavioral response measures. For one set of 15 games, participants were asked to indicate the minimum payment they would be willing to accept (WTA) to give up a bet on their answer. For the other set of 15 games, they gave a probability that would make them indifferent between a lottery with that chance of winning and a bet on their answer (*a probability equivalent*).

For half of the games, participants were asked to provide a WTA estimate after picking the winner:

For each prediction you make, the bet is to win \$15 if you're right and nothing if you're wrong. But suppose instead of the bet, we offered to pay you a bonus. What's the amount of a bonus that would make you equally happy to take the bonus or take the bet?

If you offer me \$ _____ or more, I'll take the certain payment instead of the bet.

For the other half of the games, participants were asked to indicate a probability equivalent:

For each prediction you make, the bet is to win \$15 if you're right and nothing if you're wrong. But instead of the bet, suppose you are offered to play a lottery with a payoff of \$15. What's the probability of winning \$15 that would make you equally happy to take the lottery or take the bet?

If you offer me a lottery with a _____% chance or more to win \$15, I'll play the lottery instead of the bet.

Participants were told that their bonus would be determined by selecting one prediction at random from the first set and one from the second set to determine their payoff, and that the preferences they had expressed would be honored. At the end of the procedure, participants drew cards to determine which trials would determine the payoff, and then drew card to determine the bonus amount (for WTA) or the lottery probability (for probability equivalents). The bonus amounts that could be drawn ranged from 1 to 15, and the lottery probabilities that could be drawn ranged from 50 to 100. Participants' preferences were followed in determining whether to tie the bonus to a correct prediction or instead to receive the bonus or run the lottery. For example, suppose a participant asked for a \$10 WTA and an 80% probability of winning. If the bonus number he drew was \$8 (smaller than his demand WTA), he would have to play a bet on his judgment rather than receiving the \$10 fixed prize. If the lottery probability he drew was 85% (greater than his probability equivalent), then he would get to play the lottery with a 85% chance of winning instead of betting on his own judgment.

Participants took a five-minute break after they finished the prediction tasks for the first set of the 15 games. The order of the sets and response measures was counterbalanced. At the end of the study, two bets, one for WTA, and one for the matching lottery, were selected and the final payout was determined as described in the previous paragraph.

Results

Except where noted, judgments were analyzed using a MANOVA with Trial (1-5) as a within-participants variable. Again, trial number indicates the amount of information available to the participants.

WTA. Similar to Experiment 1, accuracy was measured by the proportion of correct choice of winners. Predictive accuracy remained flat as additional information was acquired. A test for linear trend did not yield a significant result, $F(1, 19) = 0.43$, $p = 0.65$. The percentage of correct choices on the last trial, with full information, was 66% (nearly identical to the result in Experiment 1).

As predicted, confidence (measured by *WTA*) rose steadily from trial 1 to trial 5 (see Figure 6). A test for linear trend analysis on the within-participants variable, Trial, was significant, $F(1, 19) = 40.53$, $p < 0.001$. To compare accuracy and confidence, we compared the *WTA* that participants asked for with a theoretical expected value, $E(WTA)$. The later represents the *WTA* of a hypothetical set of participants who were as accurate as the real participants, but well calibrated and risk neutral. Specifically, theoretical *WTA* was derived by multiplying the average accuracy rate for each trial with the fixed payoff of the bet, \$15. The results for accuracy versus confidence are shown in Figure 6. The difference between actual and theoretical *WTA* increased linearly across trials, $F(1, 19) = 14.37$, $p = 0.001$.

Risk aversion would lead participants to give *WTA* values below the expected value, and this is the result usually observed with positive gambles like these. As shown in Figure 6, by trial 4, participants are demanding more than expected value, indicating that they have either become risk seeking or, more likely, overconfident. Heath & Tversky (1991) provide another reason why *WTAs* may be elevated: People's subjective payoff is greater from a prize won or lost by knowledge than by chance. Whether that should be considered to be part of overconfidence or a separate effect is unclear to us. In

any case, it is clear that, once again, confidence increases faster than accuracy as information accumulates.

Probability equivalents. Unlike WTA, probability equivalence judgments should not be affected by risk attitudes, since the same attitude should apply to both gambles, bets on own judgments and matching lottery. Results with probability equivalents are similar to those obtained with WTA.

As shown in Figure 7, accuracy did improve some on this task. A test for linear trend yielded a significant result, $F(1, 19) = 6.29, p = 0.021$. A Duncan multiple-range test showed that this significant difference in accuracy was due primarily to a relatively low start in trial 1. On the other hand, confidence, measured as the matching lottery probability, increased steadily as information was acquired over trials. A test for linear trend yielded a significant result, $F(1, 19) = 31.63, p < 0.001$. Comparing the two, we find an increase in overconfidence as a result of accumulation of information over time, where overconfidence is measured as the difference between accuracy and the probability equivalent. There was a significant effect of trial on the difference between accuracy and confidence, $F(4, 16) = 2.6, p < 0.04$, with a marginally significant linear trend, $F(1, 19) = 3.14, p < 0.09$.

Experiment 3

Experiments 1 and 2 demonstrate that confidence and overconfidence increase as judges gather more information sequentially. We postulate that this happens because the additional information provides cues to confidence that are not tightly linked to accuracy. Two such cues, we hypothesize, are the completeness of information and its coherence. We propose that judges measure completeness of information primarily based on the

number of items of information received, rather than by the amount of new information (i.e., added predictive values) that the additional items provide. The extent to which new cues provide added information depends on both the individual predictive values of the cues and the extent to which they are redundant with prior cues. We suspect that judges are reasonably sensitive to the individual cue values, but not always very sensitive to the effects of redundancy (Kahneman & Tversky, 1973; Soll, 1999). Thus, while additional cues tend to contribute less and less to accuracy, added cues continue to contribute steadily to apparent completeness of information, and thus confidence continues to increase.

We also hypothesize that confidence is affected by the apparent coherence of information, which hinges on the judges' ability to tie the cues into a causal explanatory structure. Often, additional information can contribute to the ease of constructing an explanatory story even when it is irrelevant to predictive accuracy (Bell & Loftus, 1989; Hastie & Pennington, 2000). We test the effects of completeness and coherence in this study by manipulating two variables.

First, relying on participants' own assessments of individual cue validity, we manipulated the order of strong versus weak cues. According to our hypotheses, strong-first participants should start with greater confidence than weak-first, because they recognize these as good cues. As weaker cues are added, confidence will level off, though perhaps not sufficiently, because these are recognizable as individually weaker cues, regardless of redundancy. Weak-first participants should start out with lower confidence, and increase more quickly as better cues come in. Here, insensitivity to cue redundancy should allow for continued, even accelerating confidence, since it will be

hard to recognize that earlier, weaker cues have stolen some of the added value of later, better cues. We also manipulated the amount of information presented in each block (either three or six cues at a time). We hypothesized that larger blocks would facilitate creation of coherent causal stories from each set of new cues, thus increasing confidence and overconfidence.

Method

Participants. Participants were 40 college students from the University of Chicago. They were knowledgeable about college football and passed the football knowledge test to take part in this experiment in exchange for \$15. In addition, a reward of \$50 was promised to the participant with the best performance. The experiment took about one hour to complete.

Task and Procedure. The tasks and procedure used in this experiment were identical to those used in Experiment 1, except for two differences in the presentation format. The first difference involved the cue presentation order, and the second involved the block size and hence the number of trials within a match for some conditions. In the strong-first condition, cues were presented approximately in descending order of perceived validity as measured in a pretest in which a group of college football fans were asked to rank order 106 football statistics by the subjective cue validity (see details in the Stimuli section of Experiment 1). In the weak-first condition, this order was reversed.

In the large-block condition, 30 cues were divided into five blocks of six cues each, as in Experiments 1 and 2. In the small-block condition, the same 30 cues were divided into ten blocks of three cues each. Because this doubled the number of judgments required from some participants, the procedure for all participants was

shortened by reducing the number of games from 15 to 9. We selected three games that were upsets, three that were close, and three that were easily predictable.

A 2 X 2 factorial design was used, with Cue Order (strong-first or weak-first) and Block Size (small or large) as between-participants variables. For comparability between block-size conditions, analyses only include data from trials 2, 4, 6, 8, and 10 of the small-block condition and trials 1 to 5 of the large-block condition. For consistency, we will refer to trials by the latter numbers, even for the small-block condition, which received half-sized blocks.

Results

Picking the winner. Our football fans were able to predict the outcomes of games with an overall accuracy of 58%, which is significantly greater than chance. Similar to Experiment 1, predictive accuracy did not improve as additional information was acquired sequentially. A test for linear trend did not yield a significant result, $F(1, 36) = 2.0, p = .16$, while a test for non-linear trend gave a significant result, $F(1, 36) = 21.4, p < 0.001$. A Duncan multiple-range test showed that this significant difference in proportion of correct choices was due primarily to a decline in accuracy at trial 2. We suspect that the dip was possibly due to an accident in the selection of matches used, as this set of nine games were not as representative as the 45 matches in Experiment 1. Comparing accuracy in trial 1 with accuracy in trials 4 and 5 showed no significant change. There was a significant cue order effect, $F(1, 36) = 4.03, p = .05$. Participants in the strong-first condition showed greater accuracy than those in the weak-first condition by 3.4%.

Confidence rose steadily from trial 1 to trial 5. A test for linear trend was significant, $F(1, 36) = 196.95, p < 0.001$, rising from 65% at trial 1 to 78% at trial 5. There were no significant main effects of cue order or block size. There was a significant interaction between Trial and Cue Order, $F(4, 33) = 15.73, p < 0.001$ (see Figure 8). Strong-first participants were more confident than the weak-first participants on early trials, but less confident at the end of the sequence of evidence and judgments. We also observed a significant interaction between Trial and Block Size, $F(4, 33) = 2.44, p = 0.05$ (see Figure 9). Confidence was distinctively low at the beginning of the task in the large block-size condition.

The discrepancy between confidence and accuracy increased markedly across trials. A test for linear trend was significant, $F(1, 36) = 33.77, p < 0.001$. A test for non-linear trend was also significant, $F(1, 36) = 18.96, p < 0.001$. We also observed a significant interaction between Trial and Cue Order, $F(4, 33) = 4.33, p = 0.002$. Overconfidence was smaller for weak-first participants at earlier trials than the strong-first participants, but was greater at later trials.

Estimating point spreads. Accuracy, MAD, was flat across trials (see Figure 10). The linear trend was not significant, $F(1, 36) = 0.22, p = 0.64$. There was no significant effect of block size, $F(1, 36) = 3.05, p = 0.09$, or cue order, $F(1, 36) = .01, p = 0.91$.

A test for linear trend on the confidence measure, MEAD, confirmed that confidence climbed from trial 1 to trial 5, $F(1, 36) = 16.54, p < 0.001$. A significant interaction of Trial and Cue Order was observed, $F(1, 36) = 4.22, p = 0.003$ (see Figure 10).

Overconfidence, $M = \text{MEAD}/\text{MAD}$, increased linearly from trial 1 to trial 5, $F(1, 36) = 20.8, p < 0.001$. There was a main effect of block size, $F(1, 36) = 3.98, p = 0.05$. Participants in the large-block condition were more overconfident than those in the small-block condition. There was also a significant interaction between Trial and Cue Order, $F(1, 36) = 3.30, p = 0.01$. As with winner prediction, strong-first participants were not more confident than the weak-first participants at trial 1, $t(1,38) = -0.83, p = 0.41$, but less confident at trial 5, $t(1,38) = -2.18, p = 0.04$.

This study replicates the basic finding that overconfidence increases with increasing information. We hypothesized that providing cues in larger blocks would facilitate creating coherent stories, and thus would increase confidence. Our results provide weak support for this hypothesis. In predicting winners, confidence increased more quickly for those receiving larger blocks of cues. In estimating point spreads, we find a marginally significant tendency for those with larger blocks of cues to be more overconfident. However, large-block participants' confidence started out distinctively low, which we did not predict. It is possible that making more judgments increases overconfidence, so that with smaller blocks, more frequent judgments counteract the effects of lower coherence.

The effects of cue order were clearer. Judges clearly did pay attention to the quality of information. In the strong-first condition, the rate of increase in judges' confidence declined as the cue validity deteriorated. On the other hand, confidence in the weak-first condition accelerated as the cue validity improved. Judges who received weak cues first ended up more overconfident in the end. One way of interpreting these results is that judges recognize that good cues render later, weaker cues useless. It is harder to

understand that weak cues can also reduce the value of later, stronger cues, if there is redundancy. Alternatively, strong-first participants impose a stopping rule to regulate their level of effort in cue integration. Once confidence reaches a certain level, judges stop bothering to try and integrate weak cues. On the other hand, the weak-first participants may invest heavily in making sense of the weaker cues (the only ones they have early on) and then treat later, stronger cues as confirming, rather than redundant, evidence.

General Discussion

The present experiments provide the clearest picture available in the scientific literature of the effects of amount of information on judgment accuracy and judgment confidence in a representative, realistically complex belief-updating judgment task. Findings from Experiments 1 and 2 confirmed our expectation that confidence and accuracy would diverge as additional information is acquired. Accuracy in predicting the winner quickly hits a ceiling, essentially after the first six cues are received, while confidence continues to rise across 24 more cues. Similarly, accuracy in estimating point spreads did not improve as information accumulated, while confidence increased consistently. Confidence judgments after the first six cues were consistent with prior research (e.g., Klayman et al., 1999): Judges were slightly overconfident (approximately 5%) in choosing between two alternatives (i.e., two competing teams), and more overconfident when giving subjective confidence intervals (about 52% and 48% of point-spread estimates were within judges' 90% intervals in Experiments 1 and 3, respectively). In both cases, overconfidence increased steadily from those levels as information accumulated. In Experiment 2, we were also able to demonstrate that the increase in

expressed confidence was also reflected in judgments and decisions that had concrete consequences for judges—in this case, determining a substantial portion of their earnings for the study.

Experiment 3 tested the effects of properties of the evidence. We distinguished apparent completeness of information from increased predictive power by varying whether cues increased or decreased in validity over the sequence. We found that judges recognized differences in the validity of individual cues, but did not fully appreciate the extent to which prior cues limited the added benefit of later, redundant cues. Judges underestimated in particular the extent to which early weak cues could reduce the added benefit of later, stronger ones. We also predicted that increasing the apparent coherence of cues would produce greater confidence. We attempted to manipulate coherence by varying the amount of information acquired in each block of informative cues. The results provide (at most) weak support for this hypothesis and we believe that block-size was not an effective operational manipulation of coherence. Although it is possible that coherence is not a major contributor to confidence in this judgment task.

Our findings corroborate earlier results showing that overconfidence increased with information (Oskamp, 1965; Peterson & Pitz, 1986; Slovic & Corrigan, 1973). Now, with this broader empirical grounding, we can go on to investigate some of the interesting implications of the observed accuracy-confidence discrepancy. For example, it is absolute or relative amount of information that drives confidence? Can judges appreciate the effects of redundant information on accuracy? Does greater experience at making a judgment generally lead to greater overconfidence? Further investigation is called for to clarify the conditions under which this will and will not be likely to happen. Whether

confidence increases with information acquisition may well depend on the kind of feedback available in the domain of judgment, the types of cues used, and the nature of the process by which information is accumulated.

The findings of the present study can also contribute to ongoing efforts to understand the processes by which people form confidence judgments, and how those judgments connect with reality. A number of investigators have contributed to emerging general theories of confidence (e.g., Gigerenzer et al., 1991; Doherty, Gettys, & Ogden, 1999; Griffin & Tversky, 1992; Juslin & Olsson, 1997.) It would be fair to say that the picture that emerges is that confidence and calibration are complex and multiply determined. This is to be expected, given that forming a confidence judgment involves acquisition, comprehension, integration, and evaluation of cues, and translation into action. The factors that moderate the accuracy of confidence judgment are likely to be similarly complex.

In the present studies, we find evidence for two main sources of divergence between confidence and accuracy. One source of divergence is judges' reliance on perceived completeness of information as a cue to confidence. The perception of completeness is largely a function of the number of cues received. Among knowledgeable judges such as those in our studies, subjective completeness may also reflect perceived differences in the predictive validity of individual cues. However, our results suggest that judges show insufficient correction for the effects of cue redundancy as information accumulates. Thus, they tend to overestimate the incremental predictive value of additional information.

Confirmation biases may be another important source of divergence between confidence and accuracy (Carlson & Russo, 2001; Hoch, 1985; Klayman, 1995; Koriat et al., 1980; Russo, Medvec, & Meloy, 1996; Schum & Martin, 1982; Sherman et al., 1983). Confirmatory comprehension and reasoning would mean that the interpretation of new evidence (statistical cues) would be biased in the direction of the previous favored hypothesis. We investigated whether confirmation bias causes confidence to increase by comparing human judges with the statistical models in their tendency to change the predictions of the winning team. For a binary choice task, confirmation bias implies greater stickiness in judgments, meaning judges would under-react to information that disconfirms their prior hypotheses. Thus, they would be less likely to change their predictions as they obtain disconfirming information.

For example, suppose the model predicts that the probabilities of winning the game for Team A are 70% and 40% at trials 1 and 2 respectively. In such a case, the model switches its predicted winner at trial 2 from Team A to Team B. Suppose a participant also predicted Team A as the winning team and indicates a 70% confidence level at trial 1. If the participant discounts the additional information at trial 2, he would either maintain his confidence level at 70% or under-adjust his confidence to, for example, 60%. In other words, he would not switch to Team B even though the cues imply that he should (i.e., we rely on the model to tell us what an objective judge would do). As a result, the overall proportion of switches a “confirmatory” human judge makes would be smaller than the model’s.

We first identified trials where the participants switched from Team A to Team B or vice versa and called these trials “switch trials”. Then we divided the number of

switch trials by the total number of trials to derive the proportion of switch trials for each individual human judge and averaged the proportion across the participants. A similar procedure was repeated to derive the proportion of switch trials for the statistical models.

Confirmatory reasoning would predict that the proportion of the switch trials would be greater for the models than for the human judges. The results across all three studies showed that human judges did not switch their predictions as often as did the models (see Figure 11). In Experiment 1, human judges changed their predictions of the winning team 6.7% of the time when additional information was acquired, whereas the models switched to the other team 11.4% of the time, $t(29) = -4.78, p < 0.001$. In Experiment 2, the proportion of switch trials were 6.7% for the human judges and 10% for models, $t(39) = -4.78, p < 0.001$. We observed similar results in Experiment 3, $t(39) = -14.86, p < 0.001$.

Moreover, we observed greater stickiness in the strong-first condition where the participants received good cues in the beginning of the task and poor cues in the end in Experiment 3. The proportion of switch trials was only 2.6% for strong-first participants, whereas it was 6.0% for the weak-first participants. Note that the proportion of switch trials for the models was greater in Experiment 3 than other studies. We suspect the difference was due to a relatively smaller sample of games used in Experiment 3 and thus the greater proportion of ambiguous games may have caused the models to change predictions at a high rate.

We have little direct evidence in our studies for another cue that we suspect is also important in confidence-accuracy divergence; the perceived coherence of the information. Research in other domains has shown that nondiagnostic information can

facilitate the formation of larger information structures (e.g., narratives, images), and thus increase the subjective coherence of the judge's mental representation of the evidence. So, even uninformative details (color of team uniforms, time of day, type of turf, weather conditions, and when the game was played – day time or evening) may increase the clarity of the judge's mental picture, and increase confidence without increasing accuracy (Bell & Loftus, 1989; Hastie & Pennington, 2000). In fact, the impression of coherence may even be partly based on redundancy, if the agreement between cues is perceived as corroboration rather than redundancy (Soll, 1999).

While conducting the third study of this paper, we collected some verbal protocols on a small number of additional judgments following the main task. These provide some interesting suggestions for further research. For example, participants rarely looked back at cues they had already considered. This implies that they were using a sequential process of updating prior estimates in the light of new information, rather than reexamining prior data. If so, this opens the door to processes that bias judgments toward prior hypotheses, such as anchoring and insufficient adjustment (Chapman & Johnson, 2002; Epley & Gilovich, 2002; Mussweiler & Strack, 2000; Tversky & Kahneman, 1974) and confirmatory distortion of information. Judges often scanned the new block of cues for especially valid cues and those that showed the largest difference between the two teams. This implicates strategies akin to those that Gigerenzer and his colleagues call “fast and frugal” (Gigerenzer & Goldstein, 1996; Gigerenzer, Todd, & the ABC Research Group, 1999), that is, strategies that allow the judge to reduce complexity by quickly focusing on the information he or she sees as most relevant. At the same time, most participants reported making inferences about general properties of the teams, such

as overall assessments of a team's "passing game," or their "global defense," based on several cues. This seemed to be most common when the teams were seen as fairly evenly matched overall, making for a difficult choice. This suggests a role for sense-making and causal thinking, especially when a decision is required in the face of conflicting information

We believe that a satisfactory analysis of the relationships between accuracy and confidence in judgment requires a multi-factor framework like the one we adopt in this paper. The key is to identify the diverse cues to confidence that judges use, and how they relate (imperfectly) to the variables that determine accuracy in a given task. Everyone studying this phenomenon would agree that the validity of information (cue validity or diagnosticity), and the differential strength of evidence for two (or more) alternative hypotheses are primary determinants of confidence, as shown in the present studies (cf., Erev, Wallsten, & Budescu, 1994; Gigerenzer et. al., 1991; Juslin & Olsson, 1997; Yates, 1990). But, our studies suggest that subjective completeness and coherence of information are also important determinants of subjective confidence. Even this is certainly not the full slate of cues to confidence. We believe that additional factors such as over-reliance on redundant and irrelevant information, the ease or fluency of the judgment experience, and the judge's personal base rate of past success making similar judgments will also play roles in a complete theory of the sources of confidence in uncertain beliefs.

References

- Babad, E. (1987). Wishful thinking and objectivity among sports fans. *Social Behavior*, 4, 231–240.
- Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investments. *Quarterly Journal of Economics*, 116, 261–292.
- Bell, B.E., & Loftus, E.F. (1989). Trivial persuasion in the courtroom: The power of (a few) minor details. *Journal of Personality and Social Psychology*, 56, 669-679.
- Carlson, K. A., & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied*, 7, 91–103.
- Chapman, G.B., & Johnson, E.J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 120-138). New York: Cambridge University Press.
- Doherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, 106, 180-209.
- Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, 85, 395–416.
- Epley, N., & Gilovich, T. (2002). Putting adjustment back in the anchoring and adjustment heuristic. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 139-149). New York: Cambridge University Press.

- Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101, 519–527.
- Ferrell, W. R. (1994). Discrete subjective probabilities and decision analysis: Elicitation, calibration and combination. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 411–451). Chichester: Wiley.
- Ferrell, W. R., & McGoey, P. J. (1980). A mode of calibration for subjective probabilities. *Organizational Behavior and Human Performance*, 26, 32–53.
- Gigerenzer, G., & Goldstein, D.G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650-669.
- Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506–528.
- Gigerenzer, G., Todd, P., & the ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
- Griffin, D., Dunning, D., & Ross, L. (1990). The role of construal processes in overconfidence predictions about the self and others. *Journal of Personality & Social Psychology*, 59, 1128–1139.
- Griffin, D.W., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411-435.
- Hammond, K (1996). *Human judgment and social policy : Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.

- Harvey, N., Koehler, D., & Ayton, P. (1997). Judgments of decision effectiveness: Actor-observer differences in overconfidence. *Organizational Behavior & Human Decision Process* (Vol. 70, 267–282).
- Hastie, R., & Pennington, N. (2000). Explanation-based decision making. In T. Connolly, H.R. Arkes, and K.R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 212-228). New York: Cambridge University Press.
- Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4, 5 –28.
- Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 719–731.
- Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, 104, 344–366.
- Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1038–1052.
- Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, 107, 384-396.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.

- Klayman, J. (1995). Varieties of confirmation bias. In J. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Psychology of learning and motivation: Vol. 32. Decision making from a cognitive perspective* (pp. 365–418). New York: Academic Press.
- Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79, 216–247.
- Koehler, D.J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 686-715). New York: Cambridge University Press.
- Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human, Learning & Memory* (Vol. 6, 107–118).
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480–498.
- Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32, 311–328.
- Larrick, R. P. (1993). Motivational factors in decision theories: The role of self-protection. *Psychological Bulletin*, 113, 440–450.
- Lichtenstein, S., & Fischhoff, B. (1981). *The effects of gender and instructions on calibration* (Tech. Rep. PTR-1092-81-7). Eugene, OR: Decision Research.
- Lundeberg, M. A., Fox, P. W., & Puncochar, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology*, 86, 114–121.

- McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probability: Theories and models 1980-94. In Wright, G. & Ayton, P. (Eds), *Subjective probability*. Oxford, England: John Wiley & Sons. (pp. 453-482).
- Mussweiler, T., & Strack, F. (2000). Numeric judgments under uncertainty: The role of knowledge in anchoring. *Journal of Experimental Social Psychology*, 36, 495-518.
- Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of consulting psychology*, 29, 261-265.
- Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles*, 48, 265-276.
- Pennington, N., Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, 51, 242-258.
- Pennington, N., & Hastie, R. (1988). Explanation-based decision making: The effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 521-533.
- Peterson, D., & Pitz, G. (1986). Effects of amount of information on predictions of uncertain quantities. *Acta Psychologica*, 61, 229-241.
- Peterson, D., & Pitz, G. (1988). Confidence, uncertainty and the use of information. *Journal of Experimental Psychology: Human, Learning & Memory*, 14, 85-92).
- Russo, J. E., Medvec, V. H., & Meloy, M. G. (1996). The distortion of information during decisions. *Organizational Behavior and Human Decision Processes*, 66, 102-110.

- Russo, J.E., & Schoemaker, P.J.H. (2002). *Winning decisions: Getting it right the first time*. New York: Doubleday Publishers.
- Schum, D., & Martin, A. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review*, 17, 105-151.
- Sherman, J., Zehner, K., & Johnson, J. (1983). Social explanation: The role of timing, set, and recall on subjective likelihood estimates. *Journal of Personality & Social Psychology*, 44, 1127–1143.
- Slovic, P., & Corrigan, B. (1973). *Behavioral problems of adhering to a decision policy*. Talk presented at The Institute for Quantitative Research in Finance, May 1, Napa, CA.
- Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, 65, 117–137.
- Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, 38, 317–346.
- Soll, J. B. & Klayman, J. (2004). Overconfidence in Interval Estimates. *Journal of experimental Psychology: Learning, memory and cognition*, 30, 299–314.
- Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard–easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, 67, 201–221.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Yates, J.F. (1990). *Judgment and decision making*. Englewood Cliffs, N J: Prentice-Hall.

Author Note

Claire I-wen Tsai, Joshua Klayman, and Reid Hastie all of the Center for Decision Research, Graduate School of Business, University of Chicago, Chicago, IL 60637; correspondence should be addressed to any of the authors. We are grateful to Paul Slovic for his generosity in sharing the reports of the empirical study of horse race handicappers that inspired the present research. We also want to thank Gary McClelland (University of Colorado), the members of several Center for Decision Research workshops who gave us useful advice on the project

Footnotes

¹The sample includes approximately 280 NCAA football conference games played at the end of the seasons in 2000-2002. The variables included were the same as the blocks of cues presented to participants in order to mimic the judgment tasks. We mimicked the judgment tasks by increasing the amount of information at the same pace as the experiments. Take Trial 1 in Experiment 1 as an example, we ran the regression using the first block of six cues that were presented to the participants. Similarly, we increased the number of variables from six to twelve at trial 2 and ran another regression model using two blocks of six cues. By the time we increased the number of variables to 30, we have built five different models for each game and each of the models includes 6, 12, 18, 24, or 30 cues. Because we had five different cue orders by counterbalancing the cues using a 5x5 Latin-Square, we have five different models for trial 1 in Experiment 1. We have 25 (5 trials x 5 cue orders) different logit regression models in total.

²We defined an upset as one in which teams with better win/loss record lost to an opponent with a weaker record. The percentage of upsets for NCAA football at the end of the seasons in 2000-02 was 17.8% in the population and 16.7% in the sample.

³Data were transformed using arcsin transformation. Data of the winner selection task in Experiments 2 and 3 were also transformed.

⁴Point spreads estimates in favor of the wrong team were coded as negative. For example, if the participant estimated Team A would win by three and Team A actually lost by six, then MAD would be nine.

⁵Data for accuracy and confidence were transformed for the analysis by using logarithmic transformation to be consistent with the analysis for overconfidence given

that the measure for overconfidence, MEAD/MAD, is proportional. Data of the point spread prediction task in Experiment 3 were also transformed.

⁶We were concerned that some participants might have mistakenly assumed that the valid range for their confidence intervals could not extend below zero (that is, could not include values that contradicted their choice of winner). Thus instead of taking the difference between the upper bound and lower bound interval estimates, we measured the confidence interval size by twice the distance between participants' estimated point spreads and the upper bound of their 90% confidence interval.

⁷The MEAD measure is based on the idea that the 90% confidence intervals participants indicated in the study represents their expectations about where the correct answer may fall. For example, suppose a participant believes that there is a 5% chance the answer will fall below the interval and a 5% chance it will fall above it, and is most likely to fall near the center of the interval. If we assume a normal distribution of probability with the given 5th and 95th fractiles, the average expected absolute deviation (EAD) between the true answer and the central best guess is, $EAD = W / (z_{.5+p/2} \times \sqrt{2\pi})$.

$W = 2 \times (\text{the upper bound of the 90\% confidence interval} - \text{estimated point spread})$.

MEAD is the average EAD across 15 games for each individual participant and can then be compared to MAD to measure over- or underconfidence.

Figure Caption

Figure 1. Accuracy of statistical models: Stepwise logit regression models were used to predict the football game winners and stepwise regression models were used to estimate point spreads. Six additional statistical cues were provided on each consecutive trial.

Figure 2. Tasks: Predicting the winner and point spread at trial 1. A match is randomly selected to illustrate the computerized football prediction tasks at the end of trial 1, showing the first block of six cues and the participant's judgments.

Figure 3. Tasks: Predicting the winner and point spread at trial 5. A match is randomly selected to illustrate the computerized football prediction tasks at the end of trial 5, showing the first block of six cues and the participant's judgments.

Figure 4. Predicted versus actual proportion of correctly chosen winners.

Figure 5. Accuracy and confidence in estimating point spreads. Accuracy is shown as the mean absolute deviation (MAD) between the true and estimated point spreads. Confidence is shown as 90% confidence intervals and the mean expected absolute deviation (MEAD) that was inferred from the width of the intervals.

Figure 6. Expected WTA vs demand WTA. WTA is the minimum amount that participants were willing to accept as payment to give up a \$15 award for a correct prediction. $E(WTA)$ is the hypothetical expected amount that would be demanded by a risk-neutral, well-calibrated player with the same average accuracy.

Figure 7. Results for probability equivalent (PE). The average PE is the minimum probability of winning a lottery that would make participants indifferent between playing the lottery and betting on a correct prediction with the same \$15 prize.

Figure 8. Predicted versus actual proportion of correctly chosen winners by cue order.

Figure 9. Predicted versus actual proportion of correctly chosen winners by block size.

Figure 10. Accuracy and confidence in estimating point spreads by cue order. Accuracy is shown as the mean absolute deviation (MAD) between the true and estimated point spreads. Confidence is shown as 90% confidence intervals and the mean expected absolute deviation (MEAD) inferred from the width of the intervals.

Figure 11. Proportion of switch trials for the models and participants for Experiments 1–3.

Figures

Figure 1. Accuracy of statistical models: Stepwise logit regression models were used to predict the football game winners and stepwise regression models were used to estimate point spreads. Six additional statistical cues were provided on each consecutive trial.

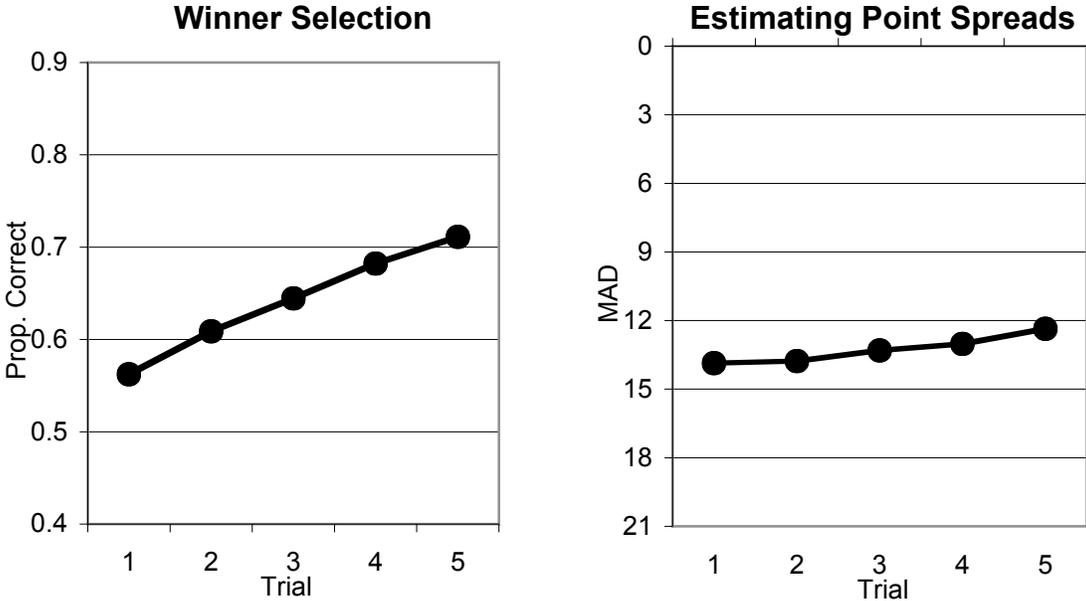


Figure 2. Tasks: Predicting the winner and point spread at trial 1. A match is randomly selected to illustrate the computerized football prediction tasks at the end of trial 1, showing the first block of six cues and the participant's judgments.

Football task

	Statistics	Team A	Team B
1. Turnover margin	Margin	8.99	8.40
2. Turnover margin	Gtotal	25.00	21.00
3. Team offense	Own fumbles (Lost)	8.00	12.00
4. Team offense	Penalties (No.)	68.00	70.00
5. Team offense	Penalties (Yds)	529.00	472.00
6. Kickoff returns	Avg (Yds/Ret)	16.31	19.00

1. Please use the football statistics to pick the winner of the game between Team A and Team B.

Team

Please indicate your confidence in your prediction of the winner in Question 1. Enter a number that best describes your confidence level.

50% 60% 70% 80% 90% 100%
not confident at all "just guessing" extremely confident "almost certain"

%

2. Please estimate the point spread for the game.

Point spread: (A beat B by) points

Indicate your confidence in your point spread estimate by giving us 90% confidence intervals (CI) around your estimate. Give us 90% bounds by writing down a pair of bracketing numbers, chosen such that you are 90% certain that the true point spread value lies between them. Your estimate upper bound for a 90% confidence interval implies that the chance for the true point spread to be greater is only 5%. Similarly, your estimate lower bound implies that the chance for the true point spread to be lower is only 5%. So, if your confidence interval is wider, it means you are less confident in the true point spread, if it is narrower, it means you are more confident in your estimate.

Move the sliders below.

Upper bound on CI: points

Lower bound on CI: points

Your prediction of point spread: 8

A lost to B by: A and B tied: A beat B by:

Click here to continue:

Figure 3. Tasks: Predicting the winner and point spread at trial 5. A match is randomly selected to illustrate the computerized football prediction tasks at the end of trial 5, showing the first block of six cues and the participant's judgments.

Statistics	Team A	Team B
1. Turnover margin	Margin	0.90 0.40
2. Turnover margin	Gtotal	25.00 21.00
3. Team offense	Own fumbles (Lost)	8.00 12.00
4. Team offense	Penalties (No.)	66.00 70.00
5. Team offense	Penalties (Yds)	529.00 472.00
6. Kickoff returns	Avg (Yds/Kret)	16.31 19.00
7. Scoring offense	Avg (Pts/#Games)	29.30 26.70
8. Scoring defense	PtsGm	27.60 24.30
9. Team offense	Firstdowns (Rush)	76.00 73.00
10. Team offense	Firstdowns (Pass)	101.00 83.00
11. Avg pts scored (against common opponents)	33.00	35.40
12. Avg pts allowed (against common opponents)	37.40	16.00
13. Overall	HomeAdv	Home Away
14. Total offense	YdsGm	330.00 318.00
15. Total offense	Avg (Yds/Play)	5.60 5.43
16. Overall	Win/Loss	6 / 4 5 / 5
17. Total defense	YdsGm	305.00 330.00
18. Total defense	Avg (Yds/Play)	5.30 6.01
19. Passing offense	YdsGm	230.00 219.00
20. Passing offense	PaPct (Pcom/Patt)	57.86 55.43
21. Passing offense	InPct	2.22 3.56
22. Passing defense	YdsGm	200.00 219.00
23. Passing defense	YdsCmp	11.23 15.34
24. Passing defense	Int	15.00 6.00
25. Rushing offense	YdsGm	110.00 135.00
26. Rushing offense	Avg (Net Yds/Carries)	3.30 2.80
27. Rushing offense	Carries	390.00 414.00
28. Rushing defense	YdsGm	89.00 96.00
29. Rushing defense	Avg (Net Yds/Carries)	3.10 2.86
30. Passing offense	Int	9.00 12.00

1. Please use the football statistics to pick the winner of the game between Team A and Team B.

Team

Please indicate your confidence in your prediction of the winner in Question 1. Enter a number that best describes your confidence level.

50% 60% 70% 80% 90% 100%
 not confident at all "just guessing" extremely confident "almost certain"

%

2. Please estimate the point spread for the game.

Point spread: (A beat B by) points

Indicate your confidence in your point spread estimate by giving us 90% confidence intervals (CI) around your estimate. Give us 90% bounds by writing down a pair of bracketing numbers, chosen such that you are 90% certain that the true point spread value lies between them. Your estimate upper bound for a 90% confidence interval implies that the chance for the true point spread to be greater is only 5%. Similarly, your estimate lower bound implies that the chance for the true point spread to be lower is only 5%. So, if your confidence interval is wider, it means you are less confident in the true point spread, if it is narrower, it means you are more confident in your estimate.

Move the sliders below.

Upper bound on CI: 15 points

Your prediction of point spread: 8

Lower bound on CI: 2 points

A lost to B by: A and B tied A beat B by:

Click here to continue:

Figure 4. Predicted versus actual proportion of correctly chosen winners.

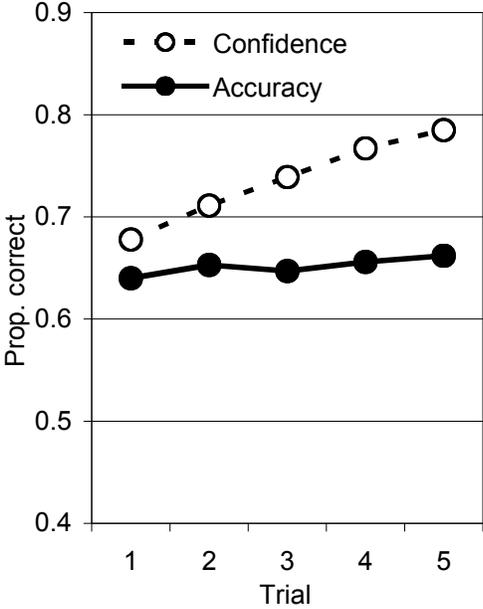


Figure 5. Accuracy and confidence in estimating point spreads. Accuracy is shown as the mean absolute deviation (MAD) between the true and estimated point spreads. Confidence is shown as 90% confidence intervals and the mean expected absolute deviation (MEAD) that was inferred from the width of the intervals.

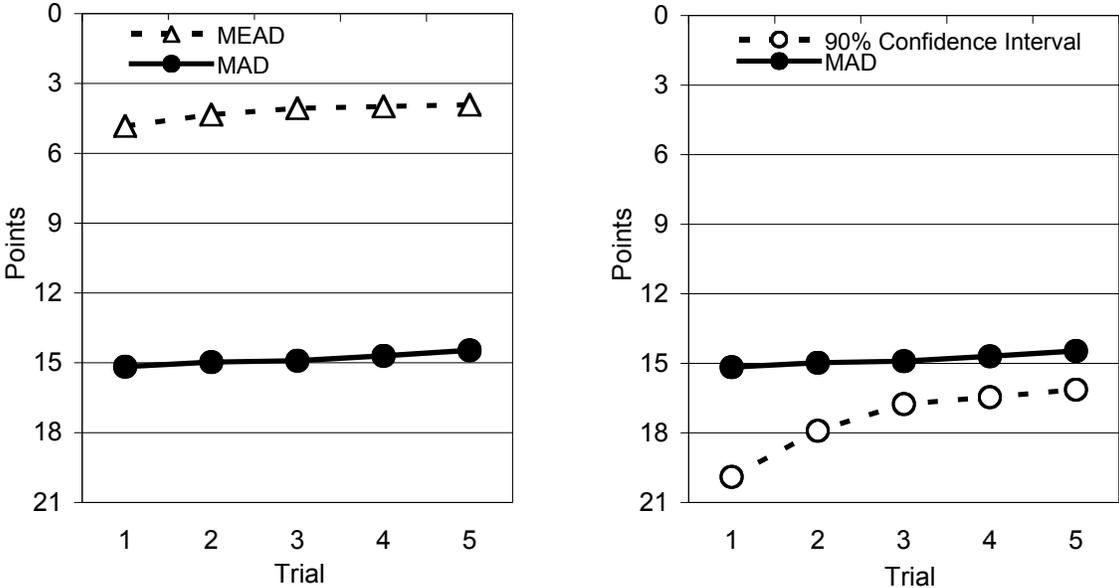


Figure 6. Expected WTA vs demand WTA. WTA is the minimum amount that participants were willing to accept as payment to give up a \$15 award for a correct prediction. E(WTA) is the hypothetical expected amount that would be demanded by a risk-neutral, well-calibrated player with the same average accuracy.

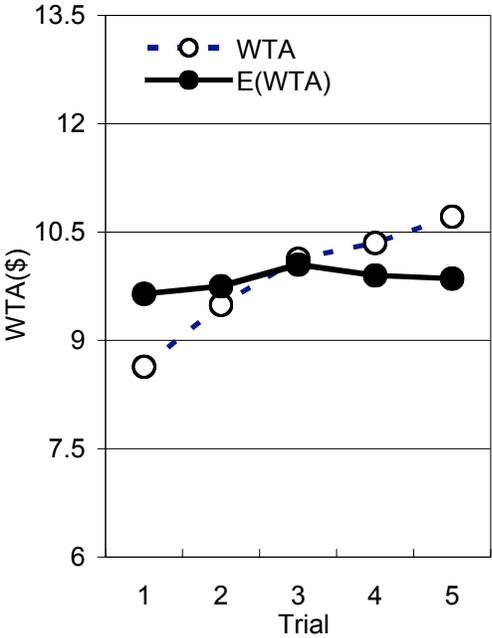


Figure 7. Results for probability equivalent (PE). The average PE is the minimum probability of winning a lottery that would make participants indifferent between playing the lottery and betting on a correct prediction with the same \$15 prize.

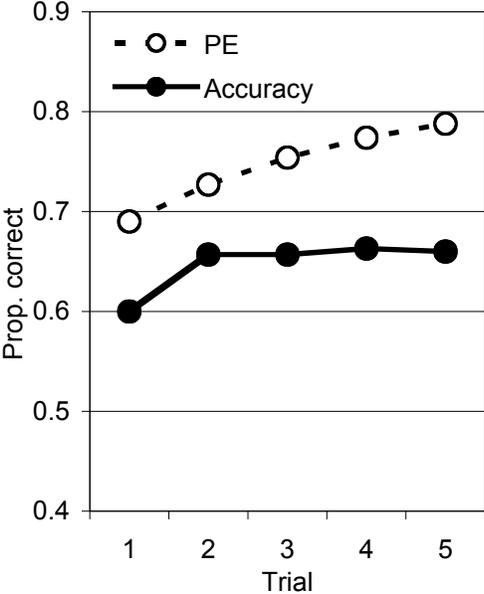


Figure 8. Predicted versus actual proportion of correctly chosen winners by cue order.

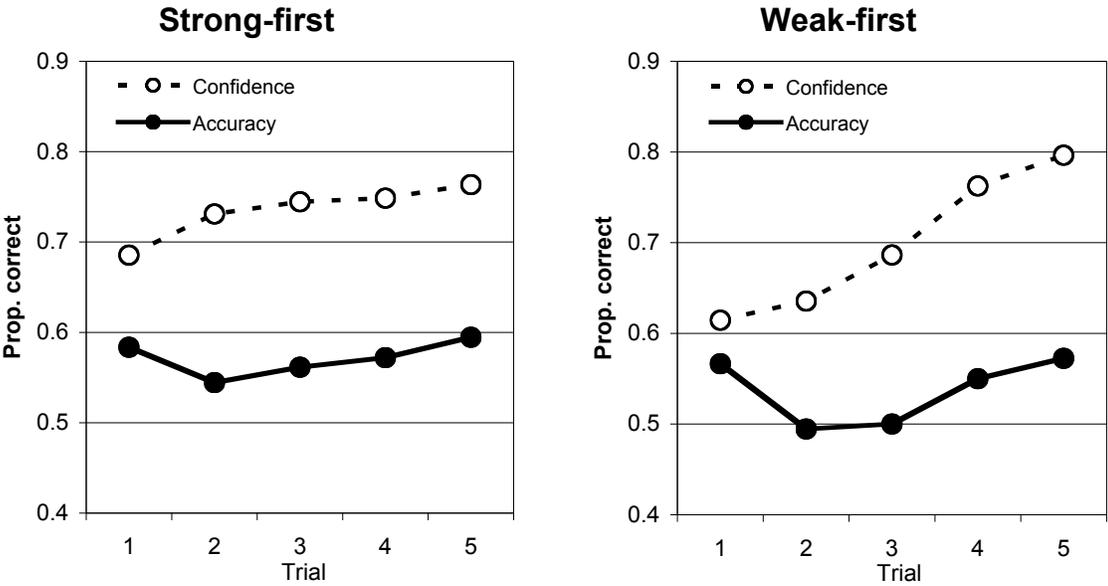


Figure 9. Predicted versus actual proportion of correctly chosen winners by block size.

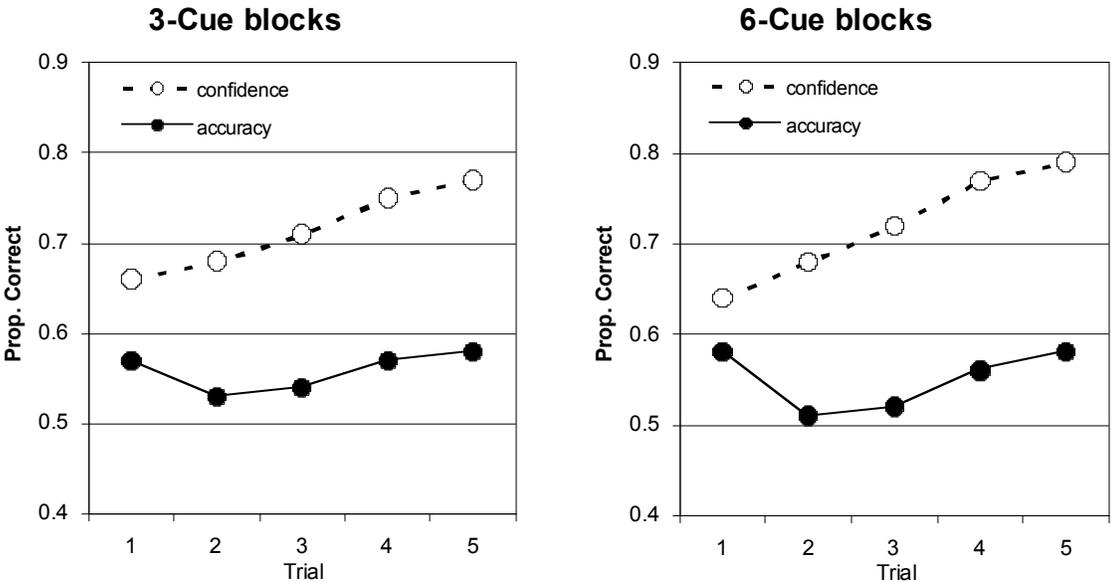


Figure 10. Accuracy and confidence in estimating point spreads by cue order. Accuracy is shown as the mean absolute deviation (MAD) between the true and estimated point spreads. Confidence is shown as 90% confidence intervals and the mean expected absolute deviation (MEAD) inferred from the width of the intervals.

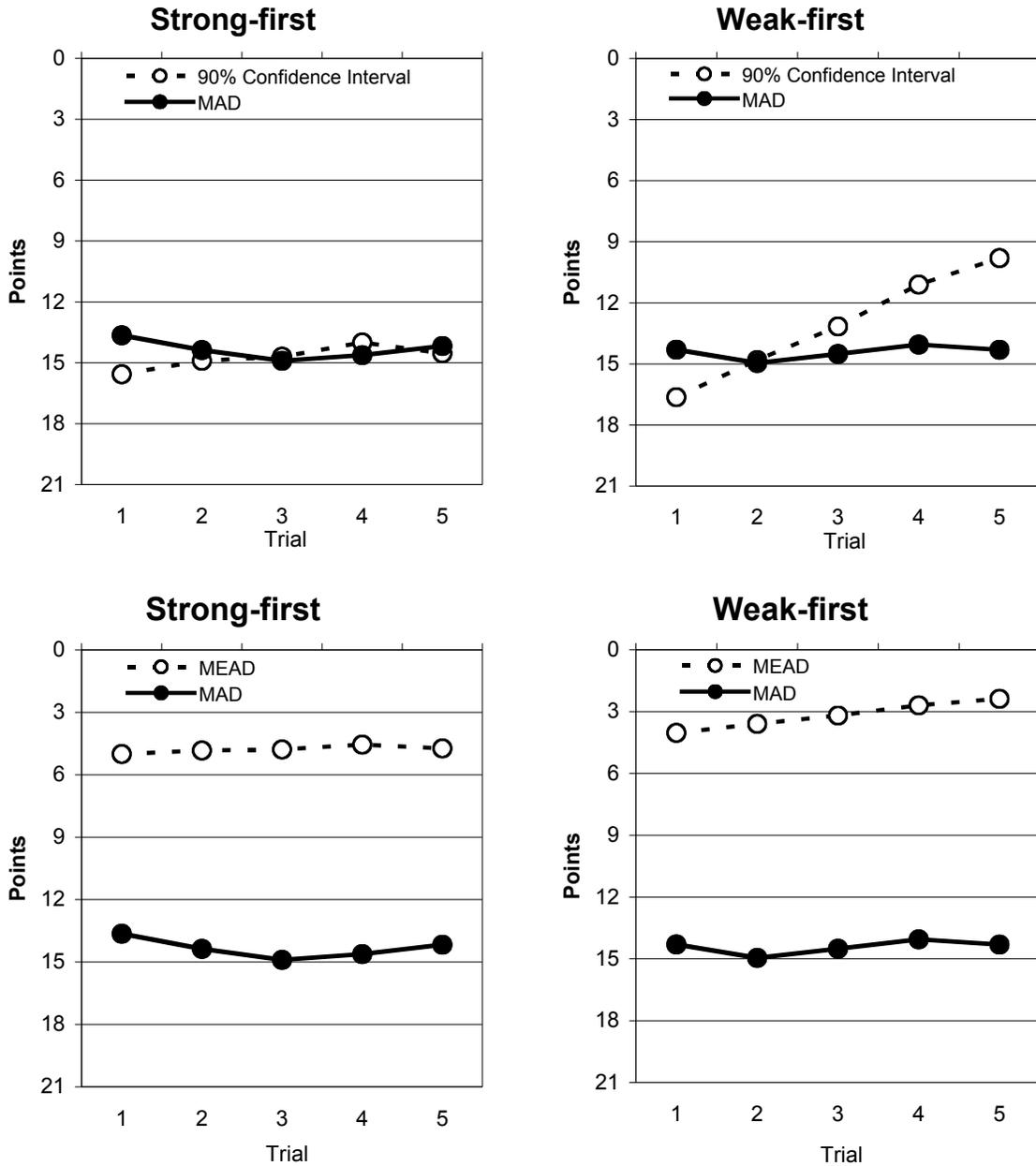


Figure 11. Proportion of switch trials for the models and participants for Experiments 1–3.

