Organizational Behavior and Human Decision Processes 107 (2008) 97–105

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

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

Effects of amount of information on judgment accuracy and confidence

Claire I. Tsai [a,*], Joshua Klayman [b], Reid Hastie [b]

[a] Rotman School of Management, University of Toronto, Toronto, Ont., Canada
[b] Graduate School of Business, University of Chicago, 5807 S. Woodlawn Avenue, Chicago IL 60637, USA

ARTICLE INFO

Article history:
Received 20 March 2006
Accepted 4 January 2008
Available online 27 May 2008

Accepted by Robyn Dawes

Keywords:
Judgment
Confidence
Accuracy
Football
Overconfidence
Calibration

ABSTRACT

When a person evaluates his or her confidence in a judgment, what is the effect of receiving more judgment-relevant information? We report three studies that show when judges receive more information, their confidence increases more than their accuracy, producing substantial confidence–accuracy discrepancies. Our results suggest that judges do not adjust for the cognitive limitations that reduce their ability to use additional information effectively. We place these findings in a more general framework of understanding the cues to confidence that judges use and how those cues relate to accuracy and calibration.

In many situations, ranging from financial investments and medical dilemmas to poker games, our degree of confidence in a judgment determines the actions we take. 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?

The focus of behavioral research on confidence judgments has been on calibration, that is, how well a judge's confidence relates to the actual probability of occurrence of the judged event or to the accuracy of a quantitative estimate. The most popular research paradigms to study 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).

Recent developments in research on confidence show that the relationship between confidence and accuracy depends on a number of different variables including elicitation format, domain, and to some extent, individual differences. It has been shown that the format of the confidence assessment has a large impact. For example, confidence in choices between two alternatives is fairly accurate whereas confidence in setting an interval around an estimate usually greatly exceeds accuracy (Juslin, Wennerholm, & Olsson, 1999; Klayman, Soll, González-Vallejo, & Barlas, 1999). Prior research also shows that some domains engender overconfidence more readily than others (Klayman et al., 1999; Soll, 1996), with some evidence that more difficult domains show greater overconfidence (Ferrell, 1994; Ferrell & McGoey, 1980; Juslin, Winman, & Olsson, 2000; Peterson & Pitz, 1988; Suantak, Bolger, & Ferrell, 1996). Researchers have also demonstrated the impact that certain methodological features can have on results. One of the most important of these features is representative selection of stimuli. In many cases confidence diverges from accuracy simply because people are unable to assess the predictive validity of judgment-relevant information perfectly (Erev, Wallsten, & Budescu, 1994; Gigerenzer, Hoffrage, & Kleinbolting, 1991; Soll, 1996). The selection of stimuli can turn imperfections into apparent biases. For example, selecting difficult questions from a domain tends to over-represent "contrary questions," that is, those whose answers happen to run counter to cues that are fairly diagnostic. This leads to overconfidence, because, unlike experimenters, judges have no way of telling that these questions are contrary (Gigerenzer et al., 1991; Klayman et al., 1999). Similarly, selecting cues that are particularly strong or particularly weak could lead to apparent biases because judges are not privy to experimenters' knowledge of relative cue strengths, and tests using particularly hard-to-pre-

* Corresponding author. Fax: +1 416 978 5433.
E-mail address: claire.tsai@rotman.utoronto.ca (C.I. Tsai).

0749-5978/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.obhdp.2008.01.005

dict domains will tend to be harder than judges think (Juslin et al., 2000).

Individual differences also matter. 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). There has, however, been relatively little investigation of the dynamics of confidence, such as how confidence changes with experience, with familiarity, or (the focus of the present paper) with the acquisition of additional information. In general, when more judgment-relevant evidence is received, accuracy and confidence should both increase, but how well do changes in confidence track changes in judgment accuracy?

Only a few prior studies have examined the relationship between changes in confidence and changes in accuracy. Oskamp (1965) found that judges (including practicing clinical psychologists) showed increasing overconfidence in their judgments when more case information was presented. In an unpublished study, Slovic and Corrigan (1973) provided horse-race handicappers with 40 different statistical cues to the performance of harness race contestants. From that set, each judge selected the specific cues he wanted to see in consecutive blocks of 5, 5, 15, and 15 cues each. Judges' confidence increased with additional information, but their accuracy did not. Peterson and Pitz (1986, 1988) had judges predict the performance of baseball teams using one, two, or three valid, non-redundant statistical cues. Confidence and accuracy both increased with the presentation of more information, but confidence increased more than accuracy did. In all of these studies, overconfidence was present from the beginning, and increased with more information. These studies suggest an interesting general tendency for more information to lead to greater overconfidence. In some of these studies, however, it is hard to tell the actual validity of the cues and the relationship between the actual and perceived cue validity. We do not know whether the judgment tasks in these studies are representative, either. Thus, it is unclear whether these characteristics to some extent lead to increases in confidence–accuracy discrepancy with additional information.

We will report three studies in which we vary amount of information available to participants, who are asked to make choices and judgments and to indicate their confidence. We postulate that confidence in a judgment is based on an ensemble of cues that are imperfectly correlated with the accuracy of that judgment. Our first two studies confirm that the amount of available information is an important one of those cues to confidence, affecting confidence more than it does accuracy. Experiment 3 shows that judges attend to more than just the sheer number of cues, however. They also respond to their impressions of how useful the cues are, though this impression again diverges from the actual effect on accuracy. We also provide evidence that the increasing confidence found in our studies is tied to the accumulation of evidence, and not just from making repeated judgments about a single event. Much more work will be needed to pin down how judges estimate confidence and how those impressions are miscalibrated. Based on our studies, we believe that one large component is judges' inability to take into account the cognitive limitations that reduce their ability to take advantage of additional information. For example, while judges seem to recognize that some cues are better than others, they may be poorly equipped to take into account redundancy with previous information (Kahneman & Tversky, 1973; Soll, 1996, 1999). Prior research in Social Judgment Theory (Brehmer & Brehmer, 1988; Cooksey, 1996; Hammond, Stewart, Brehmer, & Steinmann, 1975; Stewart, 1998) has shown that human judges tend to use only a few cues to make judgments and they usually combine these cues in a linear manner. Judges seem to overestimate their ability to use large amounts of information effectively. Thus, confidence in judgments and decisions may continue to rise long after the actual accuracy has leveled off.

The design of these studies profits from the lessons of recent confidence research. We include experiments using a representative selection of stimuli and cues in a real-world domain, namely college football. This is a domain in which events can, at least in principle, be predicted moderately well from available statistical cues. To further assure that a degree of predictive accuracy was achievable, we selected only participants who demonstrated knowledge of the domain. Because recent studies suggest that binary choice is not prone to much overconfidence, whereas interval estimates are very prone, we tested both (predicting the winner and estimating point spreads, respectively). 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. We also include confidence measures that do not require numerical expression of subjective probability, and we provide material incentives for good calibration. Because we sample from a large body of data concerning actual events, we are able to extend the analyses of earlier studies by providing normative benchmarks based on statistical modeling.

All three studies vary the amount of information within-participants, adding consecutive sets of cues to the information participants have available to them. Judgments and confidence are obtained after each block of cues is provided. We did not provide feedback during the process, in order to avoid confounding information acquisition with learning from experience during the procedure.

Experiments 1 and 2

These studies use a representative design (Gigerenzer et al., 1991), meaning that the validity of the cues and the outcomes of the events are statistically representative of the conditions that prevail in the domain as a whole. The usual way of operationalizing this is to select a sufficiently large set of events from the entire population of such events. Here, we selected college football games randomly from among all end-of-season NCAA conference football games from the 2000 to 2002 seasons, and informed participants of this. Team names were replaced with the letters A and B, to control for participants' idiosyncratic knowledge about particular teams or games beyond what was provided in the experiments. (Post testing indicated that participants could not infer the actual team names from the data.) Experiment 1 includes two types of judgment, choosing the winner and estimating the margin of victory, and asks for direct estimates of confidence (estimated probability of having chosen the winner or 90% confidence interval around the estimated point spread, respectively). Experiment 2 uses two indirect measures of confidence that are tied to monetary incentives, using methods that pit a bet on one's judgment against a well-defined bet with fixed payoff of $15.

Experiment 1

Method

Participants

Participants were 30 undergraduate 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

Each participant was asked to predict the winner and to estimate the point spread of 15 NCAA college football games given statistical information about the two competing teams. For each game, we divided 30 cues into five blocks of six cues each such that each block contained some items that judges were likely to perceive as new and useful. The order of five blocks presented was randomized. After each block, participants made predictions about the game and assessed their confidence in their predictions. Cues from prior blocks remained visible.

To assess the predictability of these tasks, we constructed regression models for each type of judgment at each trial, providing the model with the same 6, 12, 18, 24, or 30 cues available to the participants on a given trial. Logit regression was used for the binary (winner) choice. Models were derived from the approximately 280 Division I-A and Division I-AA NCAA football conference games played at the end of the 2000, 2001, and 2002 seasons. The models were built using 235 of those games and their predictive accuracy was tested on a holdout sample of 45 games. Correct selection of the winner improved steadily from 65% with six cues to 80% with 30 cues. Thus, the task of picking the winner is, in principle, relatively predictable from the cues, and more so with more information. Improvement in predicting point spreads was much more modest. The mean absolute difference (MAD) between the models' predictions and actual point spreads declined from 14.5 with 6 cues to 14.0 with 30 cues.

Design

Trial (i.e., amount of information) was a repeated measure, with six additional cues being provided on each trial. We counterbalanced the order of cue blocks between participants using a 5 × 5 Latin-Square, and the order of blocks was held constant for each individual participant. Each participant received 16 games including one warm-up game. All participants received the same warm-up game, which was excluded from analyses. The remaining 15 games were 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 games 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 2000–2002.[1] The 45 games were then divided into three subsets of 15 each, with each subset also being representative in these ways. Each subset was presented to 10 participants. The names of the teams were not provided, and 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. The 30 cues include the 10 rated as most valid, the 10 least valid, and 10 moderately valid cues (those ranked from 49th to 58th). This set of 30 cues were also used in Experiments 2 and 3. Each of the 6-cue blocks included some cues that were rated among the most valid, some rated moderately valid, and some rated least valid. The intent of this design was to provide participants with some seemingly useful information in each block of six cues, and to reduce the chance that later blocks would be perceived to be entirely redundant with previous cues.

Procedure

At the beginning of the experiment participants were told the purpose of the study was to understand their strategies to predict the outcomes of college football games. Subsequently, instructions and stimuli were presented by computer. For each game, 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 game. Information from previous blocks stayed on the screen when a new block of six cues were presented. After the presentation of each new block of six cues, 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. After participants had received all 30 cues and had made their final estimates, they started over with another game. They continued this process until they completed the prediction tasks for all 16 games.

Results

Except where noted, judgments were analyzed using multivariate analyses of variance (MANOVAs) with trial (1–5) as a within-participants variable, corresponding to the amount of information available to the participants (6, 12, 18, 24, and 30 cues, respectively). Each dependent measure was averaged across the 15 games for each participant. Hypotheses concerning changes in accuracy, confidence, and overconfidence were tested using tests of linear trends across trials.

Accuracy and confidence in predicting winners were measured as the proportion of correct predictions and the stated subjective probability of being correct. As shown in Fig. 1, accuracy remained flat as the amount of information increased, but confidence rose steadily. There was no significant trend for accuracy,[2] $F(1,28) = .78$, $p = .38$, but there were significant trends in confidence ($F_{\text{linear}}(1,28) = 43.39$, $p < .001$) and in overconfidence (i.e., the difference between confidence and accuracy ($F_{\text{linear}}(1,28) = 14.47$, $p < .001$). The 66% accuracy achieved on the last trial is significantly above chance, $t(29) = 12.81$, $p < .001$, though less than the 80% hit rate by the statistical model, $t(29) = -3.83$, $p < .001$. Confidence, on the other hand, reached 79% by the last trial.

Results were similar for point-spread estimates. Here, accuracy was measured as the absolute difference between estimated and actual point spreads and confidence was measured as the width of the 90% confidence interval the participants gave for that estimate. Accuracy was nearly flat across trials, $F_{\text{linear}}(1,28) = 3.23$,

[1] We defined an upset as one in which the team with the 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 2000–2002 was 17.8% in the population and 16.7% in the sample.

[2] An arcsin transformation was applied to winner-selection data in all three experiments.

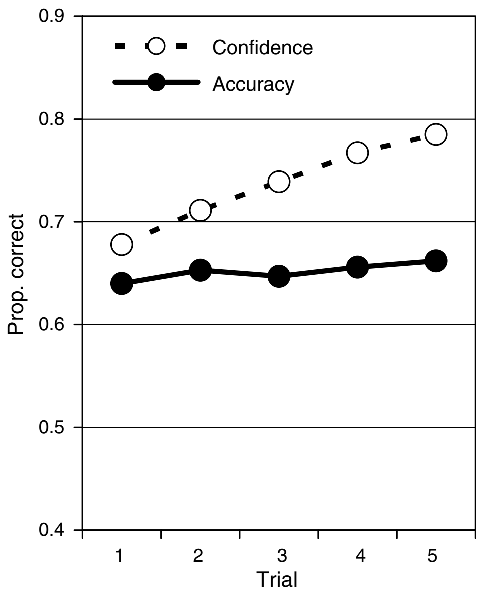

Fig. 1. Predicted versus actual proportion of correctly chosen winners in Experiment 1.

$p < .08$, whereas confidence rose steadily (i.e., confidence intervals became narrower), $F_{\text{linear}}(1,28) = 10.48$, $p = .003$. Soll and Klayman (2004) describe a measure of over- or underconfidence for confidence intervals, *M*, which compares interval widths to what would be justified given the level of accuracy (see Soll & Klayman, 2004, pp. 302–303). Using that measure, overconfidence increased significantly over trials, $F_{\text{linear}}(1,28) = 3.94$, $p = .05$.[3]

Experiment 2

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 min. In addition, a reward of $50 was promised to the participant with the best performance in this football forecast contest.

Design and procedure

The tasks and stimuli used in this experiment were identical to those used in Experiment 1, except for three modifications: (1) participants were asked to predict only the winning teams, not the point spreads; (2) two subsets of games were selected from the three sets used in Experiment 1, and participants were asked to make predictions for all 30 of those games; (3) hypothetical confidence ratings were replaced with measures that tied confidence to monetary rewards, using two variants of the Becker–DeGroot–Marschak procedure frequently followed in experimental economics (Becker, DeGroot, & Marschak, 1964).

For one set of 15 games, participants indicated on each trial the minimum payment they would be willing to accept (WTA) to give up a bet on their answer:

> 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 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*, *PE*): "...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?..." The order of the sets and response measures was counterbalanced.

At the end of the procedure, participants drew cards to determine which trials would determine the payoff (one from the set of WTA, and one from the set of PE). They then drew a card to determine a bonus amount that ranged from 1 to 15, and then another card to determine a lottery probability that ranged from 50% to 100%. For the WTA trial, if the randomly drawn bonus amount was greater than or equal to the previously stated minimum, the participant received that bonus; otherwise the participant played the bet and received $15 if the answer on that trial was correct and nothing if it was incorrect. Similarly for the PE trial, if the randomly drawn probability was greater than or equal to the stated minimum, the bonus ($15 or $0) was determined by lottery; otherwise it was determined by the correctness of the answer on that trial. An explanation of this bonus process was included in the introductory instructions for the experiment.

Results

Except where noted, judgments were analyzed using MANOVAs with trial as a within-participants variable. Similar to Experiment 1, the proportion of correct choice of winners remained flat, $F_{\text{linear}}(1,19) = .43$, $p = .65$, averaging about 65%. However, confidence, measured by WTA, rose steadily ($8.6, $9.4, $10.1, $10.4, and $10.7 on trials 1–5, respectively), $F_{\text{linear}}(1,19) = 40.53$, $p < .001$. To compare accuracy and confidence, we compared WTA to the expected value of betting on predictions, that is, $15 times the average hit rate for each trial. The difference between WTA and expected value increased linearly across trials, $F_{\text{linear}}(1,19) = 14.37$, $p = .001$.

Results with probability equivalents are similar to those obtained with WTA. Accuracy did improve some from trial 1 to trial 2 on this set of games ($F_{\text{linear}}(1,19) = 6.29$, $p = .02$) and was flat in the remaining task, but confidence, measured as the stated probability equivalent, increased steadily, $F_{\text{linear}}(1,19) = 31.63$, $p < .001$, and the difference between accuracy and the probability equivalents showed a marginally significant linear trend, F_{linear}*(1,19) = 3.14*, $p < .09$ (see Fig. 2).

Experiment 3

In this study, we examine the hypothesis that confidence is influenced not solely by the number of items of information received, but also by the perceived validity of that information. That is, confidence increases with additional information only if judges believe the additional cues are useful in distinguishing alternatives. To test the effects of perceived validity of information, we

[3] Point spreads estimates in favor of the wrong team were coded as negative. For example, if the participant estimated Team A would win by 3 and Team A actually lost by 6, then the difference was 9. We were concerned that some participants might have mistakenly assumed that their confidence intervals could not extend below zero (that is, could not include values that contradicted their choice of winner). Thus, we measured only the distance between participant's point estimate and the *upper* bound of his or her 90% confidence interval. Data on accuracy, confidence, and overconfidence in point spreads were subject to logarithmic transformation, because the last of these is a ratio.

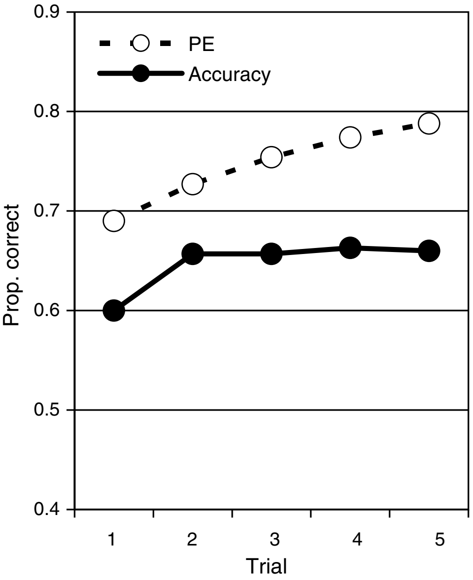

Fig. 2. Results for probability equivalents (PE) in Experiment 2. The average PE is the minimum probability of winning a lottery that would make participants indifferent between playing the lottery and betting on a correct prediction with the same $15 prize.

presented cues in either ascending or descending order of perceived validity. If judges are insensitive to cue validity, cue order should have little effect on the patterns of confidence. If validity matters, then confidence should start low and increase rapidly when starting with poor cues and adding better ones. Confidence should start out higher and increase more slowly when the best cues come early. To some extent, this difference is justified. However, we predict that insensitivity to limits in information integration will still lead to increasing overconfidence with more information. Confidence and overconfidence may increase especially rapidly when weak cues come first. Judges who mistakenly feel they can use a large number of cues at once may fail to fully jettison early, weaker information in favor of newer, better cues.

In this study, we are also able to examine another possible contribution to increasing confidence in our previous two studies. In addition to receiving more information, judges in those studies also made repeated judgments about the same event. Confidence may increase as a function of the number of times a given decision has been considered. In this study, we vary the frequency of judgments, so that participants receiving a given order of cues are asked to make judgments either after every three (i.e., 10 judgments) or after every six new cues (i.e., five judgments).

Method

Participants

Participants were 40 undergraduate students from the University of Chicago who passed the football knowledge test. They were paid $15, plus a $50 performance-based bonus that was promised to the most accurate participant. The experiment took about one hour to complete.

Design and procedure

The tasks and procedure used in this experiment were identical to those used in Experiment 1, except that we manipulated the order of cue presentation and the frequency with which judgments were made. These variables were crossed in a 2 × 2, between-subjects design. In the *strong-first* condition, cues were presented in approximately descending order of perceived validity; in the *weak-first* condition, this order was reversed. Rankings of perceived validity were taken from the pretest described in Experiment 1, in which an independent group of college football fans rank-ordered all of the available football statistics. In the *low-frequency* condition, the 30 cues were divided into five blocks of six cues each and participants made five judgments, as in Experiments 1 and 2. In the *high-frequency* condition, the same 30 cues were divided into 10 blocks of three cues each and participants made 10 judgments per game.

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. For comparability between judgment-frequency conditions, analyses use data from trials 2, 4, 6, 8, and 10 of the high-frequency condition and trials 1–5 of the low-frequency condition. For consistency, we will refer these as trials 1–5 for both conditions.

Results

Except where noted, judgments were analyzed using MONOVAs with trial as a within-participants variable and cue order and judgment frequency as between-participants variables. Each dependent measure was averaged across the 9 games for each participant. Hypotheses concerning changes in accuracy, confidence, and overconfidence were tested using tests of linear trends across trials.

Our football fans predicted the winner of games with an overall accuracy of 58%, which is significantly greater than chance. Predictive accuracy did not improve significantly with additional information, $F_{\text{linear}}(1,36) = 2.0$, $p = .16$, whereas confidence rose steadily, from 65% at trial 1–78% at trial 5, producing significant linear trends in both confidence and overconfidence ($F_{\text{linear}}(1,36) =$ 196.95 and 33.77, respectively, both $p < .001$). There appears to be a decline in accuracy between trials 1 and 2 (see Fig. 3), confirmed by significant quadratic trends in overconfidence, $F_{\text{non-linear}}(1,36) = 18.96$, $p < .001$. We suspect this dip was due only to an accident in the selection of games used, as this set of nine games were not as representative as the 45 games in Experiment 1.

Cue order had several effects (see Fig. 3) in the task of predicting winners. First, cue order had significant main effect on accuracy, $F(1,36) = 4.03$, $p = .05$; there was no significant main effect on confidence. Participants in the strong-first condition showed greater accuracy than those in the weak-first condition by 3.4% on average. Second, there was a significant interaction between trial and cue order, $F(4,33) = 43.15$, $p < .001$. As predicted, confidence rose faster in the weak-first condition. Strong-first participants were more confident in trial 1, $t(38) = -3.26$, $p = .002$, and they finished less confident than weak-first participants (though the difference was not significant, $t(38) = -1.42$, $p = .16$). This pattern was also reflected in tests of overconfidence.

Judgment frequency had little effect here. The only significant effect was an interaction between judgment frequency and trial on confidence, $F(4,33) = 2.44$, $p = .05$ Confidence at trial 1 was slightly higher in the high-frequency condition (66% vs. 64%). Results for trials 2–5 showed little difference, $F(3,34) = 1.51$, $p = .22$.

As with predicting the winner, accuracy in estimating point spreads was flat across trials, $F_{\text{linear}}(1,36) = .22$, $p = .64$, whereas confidence and overconfidence increased, $F_{\text{linear}}(1,36) = 16.54$ and 20.8, respectively, both $p < .001$. Point-spread estimates did not show the main effect of cue order on accuracy that was seen with winner selection. As before, confidence increased more rapidly in the weak-first condition, $F(1,36) = 4.22$, $p = .003$. The only effect of judgment frequency was that participants in the low-frequency condition were marginally more overconfident, $F(1,36) = 3.98$, $p = .054$.

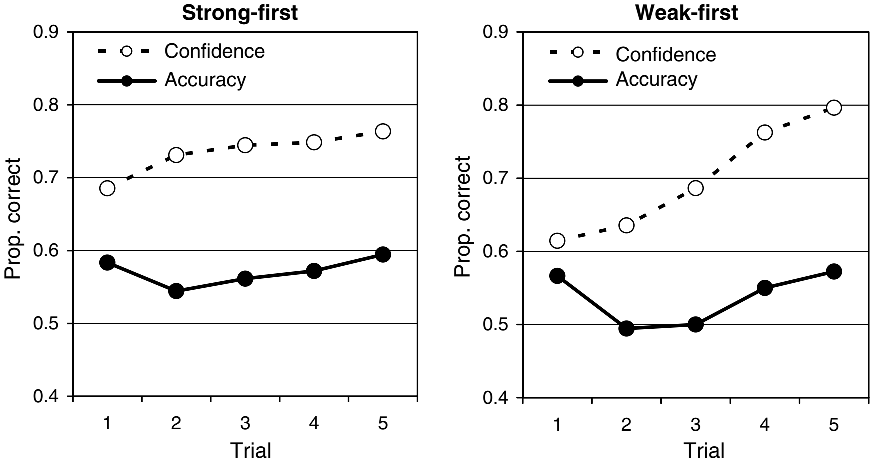

Fig. 3. Predicted versus actual proportion of correctly chosen winners in Experiment 3, by cue order.

Discussion

This study replicates the basic finding that overconfidence increases with increasing information, while also demonstrating that judges pay attention to the quality of information as well as quantity in determining their confidence. Judges who received strong cues first started out more confident than those who received weak cues first, as might be expected. When additional cues were progressively weaker, confidence increased only slowly. For those whose cues got progressively stronger, confidence increased more rapidly. These results support our hypothesis that confidence increases with additional information only to the extent that judges believe the new information is useful in distinguishing alternatives. Unfortunately, judges may in practice be unable to profit much from additional information, even if that information is useful in principle. Thus, overconfidence increases regardless.

As in Experiment 1, judges' confidence seems to track the accuracy obtained by statistical models closely.[4] For the strong-first condition, the models' accuracy rose from 71% on the first trial to 80% on the last. For the weak-first condition, accuracy rose from 67% to 80%. Our interpretation is that judges are reasonably good at perceiving the hypothetical value of the information available to them, but poor at recognizing their limitations in using that information. It is also interesting to note that participants who received weak cues first ended up more overconfident in the end. It may be the case that those who received strong cues first recognized that later, weaker cues added little to what they already had. Thus, they did not try much to integrate them into prior judgments, and accordingly did not feel they were making much further progress. On the other hand, those who received weak cues first were constantly motivated to update their predictions, and mistakenly felt that they were successfully doing so. Confirmatory biases may also play a role. 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. This possibility is supported by the overall greater accuracy in picking winners observed for strong-first participants.

Accuracy and confidence were not affected much by the frequency with which judgments were made. Confidence in picking the winner was higher on the first trial for those who had made two judgments at that point, rather than one, but when estimating point spreads, more frequent judgments led to marginally *less* overconfidence. Thus, the finding of increased overconfidence with more information does not seem to be driven in any straightforward way by the process of making repeated judgments.

General discussion

The present experiments show that providing judges with more information can lead them to become more overconfident. This possibility was suggested by a few earlier studies (Oskamp, 1965; Peterson & Pitz, 1986; Slovic & Corrigan, 1973). The present studies used a representative selection of cues and test items, tested knowledgeable judges, and included a variety of measures of confidence, including some with direct rewards for appropriate confidence. These design features provide greater evidence that the phenomenon is substantial and reliable. In the domain we studied, predictions about college football games, accuracy was significantly better than chance with the first six cues, and showed little improvement thereafter. In contrast, confidence continued to rise through 24 more cues. These results are seen both for selection of winners and for estimations of margins of victory. Prior confidence research has found that results with two-alternative choices differ greatly from those using subjective confidence intervals (Juslin et al., 1999; Klayman et al., 1999). In our studies, confidence judgments after the first six cues were consistent with prior research: 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, however, overconfidence increased steadily from those levels as information accumulated. Experiment 3 permits us to refine these findings further. We found that judges recognized differences in the validity of individual cues: It is not merely the accumulation of cues that drives confidence, but the accumulation of *seemingly useful* information.

In each of our studies, while judges' accuracy leveled off after six cues, the accuracy of statistical predictions based on the available cues continued to improve with additional information, as did judges' confidence. Indeed, judges' confidence tracked the *models'* accuracy fairly closely. It was as though their confidence levels reflected the level of accuracy that they could have achieved, if they had optimal judgment policies. These results suggest one major reason that overconfidence increases with additional information: Judges are not sufficiently aware of the cognitive limitations that keep them from profiting from large amounts of good information.

[4] We used the same method as in Experiment 1 to construct regression models, using the same cues available to each participant at each trial.

The present findings lead to a number of further questions about causes and boundary conditions. For example, what happens to confidence with the accumulation of much smaller amounts of information, when cognitive limitations are presumably less of a factor? In the high-frequency condition of Experiment 3, we can look at results after only 3 cues are presented. Participants were not overconfident at that point, although they were by the time they had 6 cues to work with, suggesting that the relationship between information and overconfidence may change quite early in the process (see also Peterson & Pitz, 1986). Another question is whether the sequential nature of information acquisition is a necessary condition for our results. Evidence from a pilot study we conducted to test the materials and procedures of Experiment 1 suggests it is not. In that pilot, 15 participants were asked to pick winners and the amount of information provided varied between participants, rather than within. Results were very similar to those obtained with the sequential presentation method used in the main experiment. For pilot participants receiving 6, 18, or 30 cues, accuracy was 64%, 65%, and 68%, respectively, and confidence was 71%, 72%, and 80%. The comparable numbers from Experiment 1 were 64%, 65%, and 66% for accuracy and 68%, 74%, and 79% for accuracy.

The amount of potentially useful information available for a given judgment is itself an important determinant of confidence and overconfidence. However, this is certainly not all that affects the sequential updating of beliefs. The belief-updating model of Hogarth and Einhorn (1992), for example, highlights that the order in which information is received can affect the final degree of confidence. Primacy or recency may predominate, depending on the length of the information sequence, the extent to which cues are mutually contradictory, whether the goal is to evaluate a hypothesis or to estimate a value, and whether judgments are made step-by-step or at the end (see also Hastie & Park, 1986). Our tasks fall into the category of long sequences, processed (to differing degrees) in a step-by-step fashion, with the goal of evaluation. In such tasks, Hogarth and Einhorn predict, and find, more primacy than recency, suggesting that revision of beliefs should decelerate as the sequence goes on. We did not generally observe this deceleration, except in the condition in which later cues were in fact less informative. This suggests that a general effect of amount of information operates on top of, or even in spite of, such order effects.

The present studies are part of a larger, ongoing effort to understand the processes that underlie confidence and its calibration with accuracy of judgments and choices (e.g., see Doherty, Gettys, & Ogden, 1999; Gigerenzer et al., 1991; Griffin & Tversky, 1992; Juslin & Olsson, 1997; Klayman, Soll, Juslin, & Winman, 2006). 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. Some of the cues used for confidence judgments are closely linked to accuracy, and if perceived accurately, they afford good calibration. These include the validity or diagnosticity of available cues and the differential strength of evidence for two (or more) alternative hypotheses (see also Erev et al., 1994; Gigerenzer et al., 1991; Juslin & Olsson, 1997; Yates, 1990). However, there are likely to be quite a few different variables that influence perception of confidence and that can produce systematic discrepancies between confidence and accuracy.

With regard to the present studies, we suspect that there are a number of different reasons why additional information can lead to increased overconfidence. Misperceptions of the effects of redundancy may play a role. If new information is largely redundant with previous information, accuracy will not improve but confidence might continue to increase, if new cues are perceived as being valid individually. Our studies do not permit a clear picture, because we deliberately spread good cues across the sequence and the overall predictive validity of the cue set increased throughout. Nonetheless, redundancy is a potentially important source of accuracy–confidence divergence that deserves further investigation (see Kahneman & Tversky, 1973; Soll, 1996, 1999).

Confirmation biases may be another important source of divergence between confidence and accuracy (Carlson & Russo, 2001; Hoch, 1985; Hogarth & Einhorn, 1992; Klayman, 1995; Koriat, Lichtenstein, & Fischhoff, 1980; Russo, Medvec, & Meloy, 1996; Schum & Martin, 1982; Sherman, Zehner, & Johnson, 1983). Confirmatory comprehension and reasoning would mean that the interpretation of new evidence 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 predicted winning team. For a binary choice task, confirmation bias implies greater "stickiness" or "conservervatism" 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 than an optimal model would.

For example, suppose the model estimates on trial T that the probability of Team A winning the game is 70%. On trial $T+1$, the model says the probability of A winning is 40%. In such a case, the model switches its predicted winner from Team A to Team B. Suppose a participant also predicts Team A as the winning team and indicates a 70% confidence level at trial T, but the participant discounts the additional information received at trial $T+1$, and only reduces his estimate of A's chances to 60%. The model prescribes a switch, but the human participant continues to endorse the current favorite.

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 proportions across the participants. A similar procedure was repeated to derive the proportion of switch trials for the statistical models.

The presence of a confirmatory bias would imply that the proportion of switch trials would be lower for the human participants than for the statistical models. As shown in Fig. 4, human judges did not switch their predictions as often as the models did, $t(29) = -4.78$, $t(39) = -\ 4.48$, and $t(39) = -14.86$ for Experiments 1, 2, and 3, respectively, all p's < .001). Across all the studies, the models switched teams 12.7% of the time compared to 5.9% for the humans. Moreover, In Experiment 3 we observed greater stickiness in the strong-first condition, in which participants received

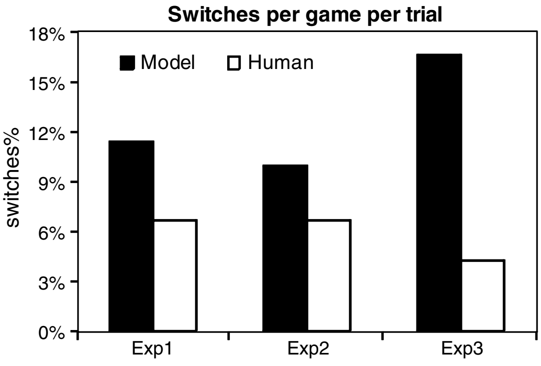

Fig. 4. Proportion of switch trials for the models and participants for Experiments 1–3.

stronger cues early in the task and weaker cues later. The proportion of switch trials was only 2.6% for strong-first participants, whereas it was 6.0% for the weak-first participants. The models switched on 13.9% and 19.4% of trials in those conditions. (The proportion of switch trials for the models was greater in Experiment 3 than other studies presumably because the sample of games used in Experiment 3 included more "close calls" than did the larger samples of games used in the other two studies.)

Another variable of interest is the perceived coherence of the information. A number of studies have shown that causal reasoning is largely influenced by the extent to which an explanation is perceived as coherent in terms of narrative structure or explanatory elements (Goldman, Graesser, & van den Broek, 1999; Graesser, Singer, & Trabasso, 1994; Hastie & Pennington, 2000; Thagard, 2000). Some information may increase coherence without increasing accuracy. For example, certain nondiagnostic details may increase the clarity of the judge's mental picture and thus increase confidence without increasing accuracy (Bell & Loftus, 1989; Pennington & Hastie, 1991). 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). We find some informal evidence of the role of sense-making and explanatory coherence in verbal protocols we collected in Experiment 3. We asked 10 participants to perform think-alouds while working on the prediction tasks. 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. Collecting additional information may have made the formation of coherent, higher-order judgments like this easier, without contributing to predictive ability.

Subjective impressions of one's mental processes have also been implicated in impressions of confidence, such as the perceived ease or difficulty of processing information (Schwarz, 2004). Furthermore, confidence may very well be influenced by other subjective experiences, such as the amount of time or effort expended to make a judgment, or the fluency with which explanations can be developed from the information given, but the impact of these meta-cognitive factors was not manipulated or assessed in the present experiments. However, we believe that a satisfactory understanding of the relationships between accuracy and confidence in judgment will require a multi-factor framework that includes both evidentiary and meta-cognitive properties of the judgment experience. 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.

Acknowledgment

We thank David Budescu, Gary McClelland, Paul Slovic, Jack Soll, Robert Zeithammer, and members of the Center for Decision Research (especially Christopher Hsee, Aparna Labroo, Ann McGill, Suresh Ramanathan, and George Wu) at the Graduate School of Business, University of Chicago for valuable comments on a draft of this article.

References

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investments. *Quarterly Journal of Economics, 116*, 261–292.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science, 9*, 226–232.

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.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT crew*. Amsterdam: North-Holland.

Carlson, K. A., & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied, 7*, 91–103.

Cooksey, R. (1996). *Judgment analysis: Theory, methods, and applications*. New York: Academic 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.

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., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

Goldman, S. R., Graesser, A. C., & van den Broek, P. (Eds.). (1999). *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso*. Mahwah, NJ: Erlbaum.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*(101), 371–395.

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. R., Stewart, T. R., Brehmer, B., & Steinmann, D. O. (1975). Social judgment theory. In M. F. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes* (pp. 271–312). New York: Academic Press.

Harvey, N., Koehler, D., & Ayton, P. (1997). Judgments of decision effectiveness: Actor–observer differences in overconfidence. *Organizational Behavior & Human Decision Process, 70*, 267–282.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or online. *Psychological Review, 93*, 258–268.

Hastie, R., & Pennington, N. (2000). Explanation-based decision making. In T. Connolly, H. R. Arkes, & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 212–228). New York: Cambridge University Press.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 719–731.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1–55.

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: Decision making from a cognitive perspective* (Vol. 32, pp. 365–418). New York: Academic Press.

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

Klayman, J., Soll, J. B., Juslin, P., & Winman, A. (2006). Subjective confidence and the sampling of knowledge. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 153–182). Cambridge, UK: University of Cambridge Press.

Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. Griffin, & 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, 6*, 107–118.

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.

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. (1991). A cognitive theory of juror decision making: The story model. *Cardozo Law Review, 13*, 519–557.

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.

Schum, D., & Martin, A. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review, 17*, 105–151.

Schwarz, N. (2004). Meta-cognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14*, 332–348.

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.

Stewart, T. R. (1988). Judgment analysis: Procedures. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 41–74). Amsterdam: North-Holland.

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.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall.