Journal of Behavioral Decision Making
J. Behav. Dec. Making, **18**: 29–53 (2005)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/bdm.486

The Recalcitrance of Overconfidence and its Contribution to Decision Aid Neglect

WINSTON R. SIECK* and HAL R. ARKES
Ohio State University, USA

ABSTRACT

Three experiments tested the hypothesis that people's overconfidence in the quality of their intuitive judgment strategies contributes to their reluctance to use helpful actuarial judgment aids. Participants engaged in a judgment task that required them to use five cues to decide whether a prospective juror favored physician-assisted suicide. Participants had the opportunity to examine the judgments of a statistical equation that correctly classified 77% of the prospective jurors. In all experiments, participants infrequently examined the equation, performed worse than the equation, and were highly overconfident. In Experiments 1 and 2, outcome feedback and calibration feedback failed to reduce overconfidence. In Experiment 3, enhanced calibration feedback reduced overconfidence and increased reliance on the equation, thus leading to improved judgment performance. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS overconfidence; judgment aids; decision aids; actuarial judgment

INTRODUCTION

Decision aids have proven to be potentially helpful in a wide variety of judgment tasks. Thanks to the advent of such decision aids, many predictions, diagnoses, and forecasts that had previously been difficult have been rendered less so. Domains in which the beneficial effects of decision aids have been demonstrated include important medical conditions (e.g., Alvarado, 1986; Corey & Merenstein, 1987), the prediction of dangerousness (e.g., McNiel & Binder, 1994), and financial auditing (e.g., Ashton, 1991).

As an example of a simple but effective decision aid, consider one developed by Corey and Merenstein (1987) designed to predict acute ischemic heart disease. The problem this decision aid was designed to solve was a quarter million unnecessary admissions to coronary care units each year for persons who had no heart disease. Using a simple regression equation using four historical variables and three electrocardiographic ones, the decision aid developed by Corey and Merenstein (1987) was highly accurate in distinguishing those

* Correspondence to: Winston R. Sieck, 1750 Commerce Center Boulevard North, Fairborn, Ohio 45324-3987, USA.
E-mail: sieck@decisionmaking.com

Contract/grant sponsor: NSF, USA; contract/grant numbers: SBR-0196200 and SES-0326468.

with ischemic heart disease from those without it. This high level of discrimination was accompanied by an impressive level of calibration ($r = +0.925$), which in this study is represented by the correlation between the aid's predicted probability of cardiac ischemia with the observed probability. Immediately following the portion of the research program when the decision aid's use was mandatory, Corey and Merenstein (1987) made the aid available for use at the discretion of the physicians in the same emergency room in which the research had initially occurred. During this period the aid was used on only 2.8% of the patients!

This abysmal level of decision aid utilization is not at all unusual. Ashton (1991) found that only 2 of 91 auditors (2.2%) consistently utilized a decision aid in a bond rating task. As was the case in the Corey and Merenstein (1987) task, the aid was a simple regression equation, this time combining three financial ratios. Graham et al. (2001) reported that 96% of emergency physicians in the United States were aware of the Ottawa Ankle Rule, which is a very simple but highly effective decision aid used in the diagnosis of ankle fractures. Nevertheless only 31% of the physicians claimed to use the rule either always or most of the time. Sixty-seven percent reported that they sometimes or never used it.

Several researchers have speculated on the reasons for the non-use of decision aids despite their demonstrated effectiveness (e.g., Dawes, 1979; Meehl, 1986; Yates, Veinott, & Patalano, 2003). Some of the proposed reasons have focused on characteristics of the aids, while others have emphasized characteristics of the end-users. As an example of the former approach, Yates et al. (2003) suggested some specific characteristics of many decision aids that may be responsible for their discouragingly low level of usage. Three features seem particularly important. The first is that many aids do not produce an output that appears to have an obviously salutary effect on the outcome of the decision. Yates et al. (2003, p. 49) point out that "... the empirical evidence for the outcome efficacy of the new procedures entailed by ... aids is statistical" in many instances. In other words, the aid may result in a correct diagnosis 87% of the time, whereas the unaided diagnostician may be correct 74% of the time. But does this specific patient represent one of those times in which the aid is correct and the physician isn't? It may be impossible to know. Therefore the decision maker is unwilling to trust the aid in this instance.

A second characteristic of aids that may discourage their use is the fact that many aids strike potential users as effortful or unnatural. The aid may require the decision maker to obtain far more or much different information than is customary. The aid may also require nothing more than a linear combination of cues, but the decision maker might strongly feel that a configural combination of cues is necessary (Hoffman, Slovic, & Rorer, 1968). Gigerenzer, Todd, and the ABC Research Group (Gigerenzer, Todd, & ABC Group, 1999) provided an instructive example of why decision makers may prefer a less-effortful unaided strategy to a complex decision aid. When trying to decide which of two German cities was larger, one possible cue a person might use is whether one of the cities has a soccer team in the major soccer league. If one city does and the other does not, the probability that the city with the team is larger than the other city is 0.87. Using more cues than the soccer cue does not give a multiple regression equation an advantage in this example. Thus a person might feel justified in using a simple intuitive strategy ("Larger cities have soccer teams, smaller cities don't") rather than a complicated one due to their equivalent efficacy.

Yates et al. (2003) also point out that many aids attempt to improve the *procedure* used in making a decision. More highly used aids aim to increase the probability of achieving a good outcome by providing information central to the *substance* of the decision, such as previously unknown information or significant new facts. The relevance of the substantive information is obvious, but the decision maker may perceive no apparent link between a new, unusual procedure and a better decision outcome.

Again, the above suggestions have emphasized properties of decision aids themselves that may contribute to their lack of acceptability. A complementary approach has focused instead on characteristics of the individual judges whom the aids are intended to benefit. For example, one hypothesized reason is the threat to one's self-concept. Auditors may think that it is their professional duty to be able to rate bonds, for instance. Abrogating one's responsibility to a simple regression equation might seem like dereliction of duty. Also, a mistaken conception of ethics might motivate the belief that using a decision aid is "dehumanizing,"

whereas using one's own intuition is somehow more caring. Meehl (1986) pointed out that sacrificing accuracy and efficiency for a warm, cuddly feeling seems like a rather shabby trade-off. Third, the belief that one is performing at a high level already might seem like a sound justification for eschewing the use of a decision aid, even if that belief is grossly inaccurate. As an example of how this might come about, Dawes (1979) has suggested that selective memory about the brilliance of one's own unaided decisions might be a factor. One might tend to recall those instances in which one was successful in making an unorthodox decision. Recollection of this victory might overwhelm the memory of those times in which one's unconventional decision was an abject failure, especially if the former is retold many times and the latter never mentioned. This performance belief, irrespective of its origins, is the contributor of focus in the current paper. That is, the current research is intended explicitly to test the hypothesis that people's overconfidence in the efficacy of their intuitive judgment strategies contributes to their reluctance to use helpful decision aids. A key underlying assumption of this overconfidence hypothesis is that usage decisions are based, at least in part, on the confidence that judges have in their intuitive strategies, as compared to their perceived performance of the aid. To the extent that usage decisions are based instead on attitudes or motivations unrelated to performance, the overconfidence hypothesis fails. Note further that basing the usage decision on confidence is unproblematic in and of itself; indeed a good case can be made that deciding based on performance beliefs is quite rational. However, to the extent that confidence is unwarrantedly high, people will fail to use helpful aids, i.e., those that outperform their intuitive strategies.

Whitecotton (1996) found evidence for the underlying assumption that the aid usage decision is based at least partly on confidence in an earnings forecasting study with professional financial analysts. Specifically, she had the analysts rate their overall confidence in their ability to perform the forecasting task after a couple of practice trials, but prior to initiating the actual task. Whitecotton found a significant negative relation between the a priori confidence ratings and reliance on the decision aid during the task. However, Whitecotton did not attempt to manipulate confidence, so the proposed causal link between confidence and aid usage is based only on correlational support.

A study by Arkes, Dawes, and Christensen (1986) provides suggestive evidence for the hypothesized role of overconfidence in decision aid neglect. These researchers first quizzed participants on their knowledge of baseball rules and regulations, and then asked the participants to indicate which of three baseball players had won the Most Valuable Player (MVP) award for each of 19 years. The researchers provided the participants with the following information for each player from the respective year: batting average, number of home runs, number of runs batted in, and the position of the player's team in the standings. They additionally provided participants with a useful decision rule; the participants were told that if they always chose the player whose team finished highest in the standings, they would get about 75% of the trials correct. The researchers found that the participants who scored more poorly on the baseball quiz relied more on the decision rule during the subsequent MVP selection task, and so significantly outperformed those participants who had done well on the quiz. The more knowledgeable group was, however, more confident in their performance on the MVP task. These results suggest that the more knowledgeable group's overconfidence reduced their reliance on the aid, and hence impaired the quality of their decisions.

Paralleling the Whitecotton study, the evidence in Arkes, Dawes, and Christensen's study for the role of overconfidence in decision aid neglect is indirect, and consists of correlational support. Hence, the causal link between overconfidence and neglect of useful decision aids is tenuous at best. For example, if the deeper reasons for aid neglect involve self-concept, dehumanization, or related attitudes and motivations that are unrelated to performance, then calibrating confidence will have little impact on aid usage. Under such a scenario, performance beliefs will be used to justify ignoring the aids only so long as such a justification is effective. Shattering such beliefs will only influence judges to find a different set of reasons that allows them to maintain their existing behavior. Such ''mental gymnastics'' have been described in the preference construction literature, as well as in research on conceptual change (Chinn & Brewer, 1993; Slovic, 1995). The primary aim of the current research is to provide direct experimental evidence for the proposal

that overconfidence contributes to decision aid neglect. Specifically, will reducing overconfidence promote greater reliance on a decision aid?

An immediate difficulty in addressing the question posed is that overconfidence itself is still relatively poorly understood, accompanied by its own long list of potential contributors (Alba & Hutchinson, 2000; McClelland & Bolger, 1994; cf. Yates, Lee, Sieck, Choi, & Price, 2002). Also, numerous distinct attempts to reduce overconfidence have been made in various task settings, and with varying degrees of theoretical motivation (e.g. Arkes, Christensen, Lai, & Blumer, 1987; Griffin & Buehler, 1999; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1980; Sieck & Yates, 2001). However, several of the proposed manipulations have met with only mixed success. For example, Koriat et al. (1980) hypothesized that overconfidence stems at least in part from an inclination to rely more heavily on reasons supporting a chosen answer than on reasons contradicting it. In order to test this proposal, they had participants in an experimental condition write reasons for and against each of a pair of alternatives given in a general knowledge test, prior to rendering judgments. Consistent with their proposal, participants in the experimental group were less overconfident than those in a control group. However, Fischhoff and MacGregor (1982) did not find the beneficial effect in a forecasting task, and there have since been unknown numbers of unpublished studies in which the results failed to replicate the basic result (Yates, Lee, & Shinotsuka, 1992). Such an outcome is not especially unusual in this literature, and likely reflects the complexity associated with the phenomenon itself. Hence, a secondary objective of the current study is to provide evidence on the efficacy of specific methods for reducing overconfidence and, hopefully, shed some light on their associated theoretical mechanisms along the way.

In order to test the hypothesis that overconfidence contributes to decision aid neglect, we consider three methods for reducing overconfidence. The methods all entail various kinds of feedback manipulations, and are supported by prior empirical findings and/or theoretical considerations. The first method considered for reducing overconfidence is to provide regular outcome feedback during an initial ''training'' segment of the task. The second supplements such outcome feedback with instructions to attend especially to cases that are likely to yield negative feedback. In the third method, aggregated calibration feedback is presented at the end of a training segment. These methods are discussed in greater depth in the introductions to the corresponding experiments. The general approach taken in these experiments is as follows. Participants engage in two phases of a classification task with real-world stimuli. Specifically, they classify general social survey (GSS) respondents as either favoring or opposing legal suicide for the terminally ill. They also report probabilities on a 50–100% scale that their classifications are correct. Either during the first phase of the task, or at least prior to initiating the second phase, participants in experimental conditions receive some form of feedback intended to reduce their overconfidence. Baseline control participants do not receive any feedback. Some participants also have the option to examine the output of a statistical decision aid at this stage, depending on experiment and condition. Then, during the second phase, participants perform the task again, without any feedback but generally with the aid available. Categorical and probabilistic judgment accuracy (including overconfidence) are assessed, along with measures of reliance on the statistical decision aid. A clear failure of the hypothesis that overconfidence contributes to decision aid neglect is found if an experimental manipulation reduces overconfidence, but is not associated with increased reliance on the aid. Reductions in overconfidence that are accompanied by greater reliance on the aid tend to support the hypothesis.

EXPERIMENT 1: BASELINE USAGE AND OUTCOME FEEDBACK EFFECTS

Several researchers have proposed that experience that includes regular outcome feedback in judgment tasks with highly related events ought to lead to good calibration (e.g., Dougherty, Gettys, & Ogden, 1999; Keren, 1987). For example, Keren (1987) had bridge experts and less experienced players judge their chances that they would make their contracts during play. He found that the experts were extremely well calibrated, whereas the

less experienced players were overconfident. Keren attributed his findings to the fact that the experts had far more practice in this highly structured task that provides regular outcome feedback. Dougherty (2001) has claimed that overconfidence should decrease as a function of experience that includes outcomes based on his Minerva-DM (MDM) memory model of judgment. In order to test the claim, he had study participants provide probabilistic differential diagnoses in a category learning task (cf. Yates, Lee, Shinotsuka, Patalano, & Sieck, 1998). The study consisted of two phases. During the initial study phase, participants examined symptoms and disease outcomes of either 80 or 240 fictitious patients, one at a time. Memory for the just-studied patients was occasionally tested, but no judgments were rendered during this phase. In a subsequent testing phase, all participants judged disease and probability for 48 ''new'' patients. No feedback was provided during this phase. Dougherty found that the group who had experienced more patients and disease outcomes achieved a higher proportion of correct diagnoses, and virtually equivalent confidence compared to the less experienced group. Thus, calibration was improved in the ''high'' experience condition.

Other researchers are pessimistic that reasonable experience, even including feedback, can have a salutary effect on reducing overconfidence. For example, Hammond, Summers, and Deane (1973) showed that in a multiple-cue probability learning task, providing outcome feedback (the correct answer) actually resulted in lower performance compared to a group given no outcome feedback. Brehmer (1980) expressed pessimism that in probabilistic tasks the provision of outcome feedback would result in substantial improvement in performance. Assuming that people become more confident in their performance as their experience on a task increases (Einhorn & Hogarth, 1978), then very sluggish performance increases combined with marked confidence increases over trials must eventually result in overconfidence. Arkes et al. (1986) suggested that providing outcome feedback on such tasks causes decision makers to be more likely to switch hypotheses after incorrect trials (cf. Levin, 1975). However, because the relation between the cues and the true outcome is probabilistic, decision makers might not know whether their change in strategy may, in fact, result in their abandonment of the optimal strategy (see also Peterson & Pitz, 1986). Finally, based on an exemplar retrieval model (ERM), Sieck (2003) argued that experience and regular feedback would not be sufficient to yield good calibration. According to the ERM, people learn by storing example cases, and base their judgments on the first few examplars that they happen to retrieve. Hence, the samples of cases that underlie judgments are too small to yield good calibration, irrespective of the number contained in a long-term store (LTS).

Thus the theoretical and empirical results presented above are equivocal as to whether regular outcome feedback reduces overconfidence. If indeed it does, and if overconfidence contributes to decision aid neglect, then providing outcome feedback should reduce overconfidence and thereby increase reliance on the decision aid. This line of thought is tested in the first experiment by manipulating the presence of outcome feedback during the initial phase. In addition, to ascertain whether the decision aid actually offers some measure of improved performance in the task, availability of the aid was also manipulated.

Method

Participants

Study participants were 114 undergraduate students enrolled in an introductory psychology course at Ohio State University. Experimental participation was part of their course requirement.

Probability accuracy measures

The most widely used method for assessing the accuracy of a set of probability judgments is the mean probability score ($\overline{PS}$), sometimes called the Brier score, given by:

$$\overline{PS} = \frac{1}{N}\sum_{i=1}^{N}(f_i - d_i)^2, \tag{1}$$

where f_i is the probability judgment for the ith case, d_i is the outcome index for that case (1 if target event occurs; 0 if not), and N is the number of judgments. A number of decompositions of the Brier score have been developed that yield measures of specific components of accuracy (for reviews, see Yates, 1990, 1994). Here, we adopt the ''covariance decomposition'' developed by Yates (1982). This approach has gained fairly wide popularity in a number of disciplines (e.g. Arkes et al., 1995; Price, 1998; Wilkie & Pollock, 1996). This is partly because it does not require any categorization or other transformations of the probability judgments, but also because the component measures are readily calculated and easily understood as the familiar estimable components of a linear model (though the calculations do not depend on parametric assumptions associated with such models). The component accuracy measures resulting from the covariance decomposition are:

$$bias = \bar{f} - \bar{d}, \tag{2}$$

$$slope = \bar{f}_1 - \bar{f}_0, \text{ and} \tag{3}$$

$$scatter = \frac{1}{N}[N_1 Var(f_1) + N_0 Var(f_0)]. \tag{4}$$

In words, bias is the mean of the probability judgments ($\bar{f}$) minus the mean of the outcome index ($\bar{d}$), slope is the mean of the judgments when the outcome occurred ($\bar{f}_1$) minus the mean when the outcome did not occur ($\bar{f}_0$), and scatter is the weighted average of the variances of the judgments, conditional upon the outcome's occurrence. Interpretation of these component measures depends on the nature of the target event. Specifically, target events can be specified in a manner external to the judge (e.g., ''It will rain'') or internal to the judge (e.g., ''My categorical prediction that it will or will not rain is correct'') (cf. Yates et al., 1998). Since the focus of the current study is on the relationship between judges' confidence in their categorical predictions, and the willingness of those judges to rely on statistical judgment aids, the internal target is appropriate for the current study. Hence, the outcome index's target event is ''My categorical judgment is correct,'' and the probability judgment is ''The probability that my categorical judgment is correct.'' Given these specifications, *bias* can be seen to measure over-/underconfidence in categorical accuracy, with positive bias indicating overconfidence and negative bias indicating underconfidence. *Slope* measures the amount of separation between the mean probability given that the categorical judgment is correct and the mean probability when the categorical judgment is incorrect. It can be loosely thought of as indicating the respondent's metacognitive assessment about whether they are right or just guessing on individual items. Finally, *scatter* is the amount of probability judgment variation that is unrelated to whether the categorical judgment is correct or incorrect, i.e., it measures the error variance.

In addition to these measures, one can also examine the correlation between the outcomes and probability judgments, as is common in research on feeling-of-knowing (Nelson, 1984). Here we used Pearson's product–moment correlation between correctness and probability judgments because of its close correspondence with slope and scatter.

Cover story and statistical equation

The cover story was adapted from that of Yates and Estin (1996). Study participants were asked to imagine that they were attorneys in the following scenario:

> In this experiment, you are to assume the role of a new attorney in a law firm, Brown & Black. Brown & Black does a lot of jury trial work. Thus, an important skill every attorney must have or acquire is the ability to anticipate how potential jurors would feel about a given issue. The supervising partner to whom

you report, Philip Elkin, has prepared a test of your juror judgment skills. If Mr Elkin concludes that your skills are just too weak, he will have to let you go. All the prospective jurors in the given jurisdiction completed a general questionnaire concerning several personal characteristics as well as their opinions about various miscellaneous issues. Those prospective jurors also responded ''yes'' or ''no'' to the question of whether a terminally ill patient should be allowed to end his or her own life. Your job is to decide how the jurors responded.

For convenience in description, a ''yes'' response by the juror is arbitrarily labeled the *target* or *focal event*, and the cues presented in the juror profiles are abbreviated as follows: A = age in years, P = political party affiliation (0 = strong democrat to 6 = strong republican), C = alcohol consumption (1 = user; 2 = total abstainer), R = religious service attendance (0 = never; 8 = several times a week), and S = belief about how wrong it is for a man and a woman to have sexual relations before marriage (1 = always wrong; 4 = not not wrong at all).

Data set and statistical equation. Data for the study were taken from the GSS database for 1993. The year was chosen because it had many cases with complete data on all variables. There were 470 cases total with complete data, and these were randomly split into a set of 286 that was used to fit the model, and a hold-out set of 184 that was used in the experiment proper. The overall base rate for the target event was 61.7%, and this did not vary substantially in the two subsets (63% and 60%, respectively). The statistical equation was created by way of a logistic regression analysis. The model was

$$p = \frac{e^Y}{1 + e^Y}, \tag{5}$$

where $p = P^*(T)$ was the model's probability judgment for the target event (in this case, a ''yes'' response to the suicide question). Y was the following linear combination of variables:

$$Y = 0.31 - 0.02\mathrm{A} - 0.26\mathrm{P} - 0.37\mathrm{C} + 0.21\mathrm{R} + 0.63\mathrm{S} + 0.30\mathrm{P}^*\mathrm{C} - 0.27\mathrm{C}^*\mathrm{R} \tag{6}$$

The proportion of variance explained by the model was $R^2_{\mathrm{L}} = 0.24$. Also, the model's judgments achieved 77% correct predictions on this training data set. Cross-validation performance results for the equation, each individual cue, and judgments based on the training/historical base rate are presented in Table 1. As can be seen, the model achieved both a higher percentage of correct deterministic predictions and a lower (improved) $\overline{PS}$ than any of the individual cues or the historical base rate. Any cue and the historical base

Table 1. Accuracy of historical base rate, individual, cue, and equation, ''judges''

Judge	Proportion correct	$\overline{PS}$
HBR	0.60	0.2414
A	0.65	0.2245
P	0.60	0.2423
C	0.69	0.2165
R	0.64	0.2309
S	0.70	0.2107
Equation	0.74	0.1964

Note: HBR = historical base rate; A = age; P = political party affiliation; C = alcohol consumption; R = religious service attendance; S = attitude towards premarital sex.

rate would improve over the uniform judgment strategy of reporting 50% on every trial. For reference, that strategy would achieve a mean probability score of 0.25.

Design

The experiment was a 2 (feedback) × 2 (equation access) between-subjects design. For each juror in Block 1, participants in the feedback condition learned what position the juror actually held, after they had finished recording their judgments. No-feedback condition participants received no such information. Participants in the equation access condition had the opportunity to view the statistical equation's choice and probability correct (i.e., $P^*(T)$ if target ''chosen''; $1 - P^*(T)$ otherwise). They were also given the following additional instructions regarding the equation and how it should be used:

> In addition to having the prospective juror's profile to guide your judgments, you will also have the opportunity to examine the prediction obtained from a statistical equation. Mr Elkin hired a statistician to derive the equation, because he thought it might help in making these sorts of predictions. The statistician used the responses from a similar group of people who responded to the same survey in order to build the equation. The statistician reported that, for the group of people whose responses were used to build the equation, the equation correctly identified viewpoints on assisted suicide 77% of the time. However, the statistician also cautioned that the equation's performance may be slightly lower when the equation is used to predict the viewpoints of new people who were not used to build the equation, such as will be the case here. Hence, 77% simply offers a rough guide as to how well the equation might work.
>
> On the following screens, you will have the option to click a button to display the prediction from the resulting statistical equation. The prediction will include the equation's ''opinion'' on whether each potential juror favors or does not favor the suicide option, as well as the equation's probability that the opinion is correct. If you do not wish to see what the equation predicts regarding a particular juror, then you do not have to click the button. It is completely up to you as to how and how often you use the equation. You may follow the equation exactly for every juror, completely ignore it, or something in between. You should do whatever it is that you think will lead you to achieve your best performance on the task.

Participants in no-equation access conditions were not informed that any such equation existed, and did not have any visible button to click.

Procedure

The experiment was conducted using computers. The program first introduced the scenario and initial instructions to the participant. The initial instructions truthfully informed participants that the ''jurors'' were randomly selected respondents from a real survey. They also stated that, for each prospective juror, participants were to indicate whether they thought the juror responded ''yes'' or ''no'' to the suicide question, and then report their confidence in the form of a 50% to 100% probability that their selection was correct. Some reasons for the importance of these probability judgments in the context of the cover story were then provided.

Participants were also given these specific instructions concerning use of the probability scale:

(1) 50% should mean that the prospective juror is just as likely to favor as to oppose the suicide option.
(2) 100% should mean that the prospective juror's position is absolutely certain to be as you indicated.
(3) Increasing probabilities between 50% and 100% should correspond to increasing degrees of certainty that the juror's actual position on suicide is as you stated.

(4) Your probability judgments should correspond to the percentage of the time that you expect to be right over all cases for which you have that same level of confidence. For example, over the times that you say ''80%,'' you should expect to be correct about 8 times out of 10. They were further told: ''Note that you should never report a probability lower than 50% in this task. If you do, you are effectively saying that you think your chosen answer is less likely to be correct than the other alternative. If you really feel that way, then you should select the other alternative and report your probability that it is correct. For this reason, you will only be given the option to report probabilities of at least 50%.''

Participants were given four practice trials with prompts to ask questions to ensure familiarity and understanding of the basic task and probability scale. Participants then rendered judgments for each of 120 prospective jurors during an initial block of trials, and after a short rest break, judged another 60 during a second, test block. Feedback was provided to feedback condition respondents only during Block 1. The statistical equation was available to participants in the pertinent conditions for both blocks. Analyses focus on Block 2, the final test, in all experiments, where no feedback is given in any condition.

After participants completed Block 2, they were asked to respond to some post-session questions regarding aggregate confidence in the just-completed task. We recognize that one's retrospective aggregate confidence following a completed series of trials often differs from the confidence expressed contemporaneously during the trial-by-trial behavior (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Sniezek, Paese, & Switzer, 1990). The trial-by-trial confidence is often too high, whereas the retrospective aggregate confidence level is lower and often too low. We also recognize that the more accurate retrospectively assessed confidence represents something of a hollow victory in that this better calibrated confidence level is expressed too late; all the trials have been completed. Any modesty expressed at this point can no longer foster the use of a decision aid, because the task has ended. Therefore our efforts are primarily designed to improve the calibration of the trial-by-trial estimates. Nevertheless, we wanted to ask retrospective questions in order to gain additional insight into the relation between the participants' decision aid usage and confidence.

First, participants in the equation access conditions were asked to reflect back on the last set of cases that they had considered (Block 2), and to consider all of the times that they felt their reported choice matched that of the equation. They were then asked to estimate their percentage of correct choices in these circumstances on a scale from 0% to 100%. The instructions noted that a person who was completely guessing should get about 50% correct. Next, these participants were asked to reflect back on all of the times that they felt their reported choice ran counter to that of the equation. They were asked to estimate both the percentage correct for themselves and for the equation in these circumstances. The instructions pointed out that in these cases, either the participant was correct or the equation was correct, so that the two estimates must sum to 100% (and the program enforced this). The pointed nature of these instructions was intended to generate a strong assessment of respondents' confidence in the accuracy of their intuitive judgment.

Finally, participants in all conditions completed the ''Attitude Toward Field'' subscale of the ''Attitudes Toward Statistics'' (ATS) scale (Schultz & Koshino, 1998; Wise, 1985). We wanted to examine the hypothesis that one's attitude toward statistics might be responsible for one's willingness to rely on a statistically-based decision aid.

Results

Probability accuracy measures

Table 2 shows the means and standard deviations of $\overline{PS}$, confidence, proportion correct, bias, slope, scatter, and Pearson's *r* for each condition, and Table 3 shows the corresponding inferential statistics. As can be seen, equation access resulted in significantly higher proportion correct, slope, and *r*, and a lower (better) $\overline{PS}$. None of the other effects were statistically significant. In particular, receiving feedback in Block 1 did not result in any judgment improvement on the final test.

Table 2. Means (SDs) for confidence and accuracy, Experiment 1

	Feedback	
	No	Yes
No equation access		
N	29	26
$\overline{PS}$	0.265 (0.044)	0.258 (0.050)
Confidence	0.76 (0.116)	0.75 (0.107)
Proportion correct	0.61 (0.076)	0.62 (0.075)
Bias	0.15 (0.107)	0.13 (0.119)
Slope	0.03 (0.034)	0.04 (0.050)
Scatter	0.013 (0.009)	0.015 (0.009)
Pearson's r	0.15 (0.118)	0.13 (0.156)
Equation access		
N	29	26
$\overline{PS}$	0.227 (0.043)	0.236 (0.041)
Confidence	0.77 (0.041)	0.76 (0.076)
Proportion correct	0.65 (0.082)	0.65 (0.075)
Bias	0.12 (0.081)	0.11 (0.070)
Slope	0.07 (0.031)	0.04 (0.045)
Scatter	0.015 (0.007)	0.015 (0.007)
Pearson's r	0.25 (0.127)	0.17 (0.162)

Table 3. F-statistics for confidence and accuracy, Experiment 1

	Factor		
Dependent variable	Equation	Feedback	Interaction
$\overline{PS}$	13.19**	0.04	0.84
Confidence	0.01	0.52	0.02
Proportion correct	4.52*	0.03	0.19
Bias	2.61	0.68	0.25
Slope	7.77**	1.53	3.12
Scatter	0.48	0.33	0.13
Pearson's r	8.30**	3.29	1.42

*$p < 0.05$; **$p < 0.01$.

Equation inspection and congruence

For participants in the equation access conditions, button clicks indicated whether the equation output was examined on each trial. Participants in the feedback condition clicked the equation button on only 30% of the trials on average, and those in the no feedback condition clicked an average of 43% of the time. The differences were not statistically significant. In general, participants' intuitions are fairly congruent with the equation, with participants' choices in the no equation access conditions matching the equation on 75% of the trials. This figure was 80% for participants who had access to the equation, a significant gain, $F(1, 110) = 4.02$, $p = 0.047$. These ''equation congruence'' figures did not differ by feedback condition.

Post-test questions

Retrospective aggregate confidence and overconfidence. As discussed above, following completion of the judgment task, participants in the equation access conditions reported two aggregate confidence estimates,

Table 4. Means (SDs) for aggregate confidence and bias (overconfidence), Experiment 1

	Feedback	
	No	Yes
Choice congruent with equation		
Agg. confidence	0.77 (0.202)	0.66 (0.231)
Agg. bias	0.03 (0.215)	0.10 (0.242)
Choice incongruent with equation		
Agg. confidence	0.61 (0.195)	0.49 (0.214)
Agg. bias	0.37 (0.318)	0.24 (0.270)

one of which pertained to their confidence when they believed that their choices were congruent with the equation, and the other of which dealt with confidence when choice and equation were incongruent. Furthermore, aggregate bias or overconfidence was computed by subtracting each participant's actual proportion correct under the specified conditions from that person's respective aggregate confidence estimate. Recall that the aggregate confidence estimate was elicited on a 0–100% scale. The means and standard deviations of the confidence and bias values are presented in Table 4. As can be seen, aggregate confidence and overconfidence are reduced significantly for the feedback over the no-feedback group for the ''choice-congruent'' estimate, $t(56) = 2.01$, $p = 0.049$, and $t(56) = 2.18$, $p = 0.034$, respectively. And they also appear to be reduced for the ''choice-incongruent'' estimate, though the overconfidence result is marginal, $t(57) = 2.24$, $p = 0.029$, and $t(57) = 1.84$, $p = 0.071$, respectively. The most striking results, however, are the levels of confidence and overconfidence associated with the choice-incongruent estimate. In particular, overconfidence is massive and highly significant for both the feedback ($+0.24$) and no-feedback ($+0.37$) groups, $t(28) = 4.83$, $p < 0.001$, and $t(29) = 6.71$, $p < 0.001$, respectively. In sum, even when making global assessments under conditions that highlight that only intuition or the analytic equation and can be correct, respondents estimated that intuition yielded correct responses about half of the time, and those estimates were much too rosy.

Attitudes towards statistics. Across all conditions, scores on the ATS ranged from 49 to 95, with an overall mean of 75.11. There were no significant differences on ATS by condition. For the equation access conditions, the correlations between ATS score and equation inspection and congruence values were positive and reasonably strong, $r = 0.54$ [$t(57) = 4.85$, $p < 0.001$], and $r = 0.52$ [$t(57) = 4.56$, $p < 0.001$], respectively. In addition, for the no-equation conditions, the correlation between ATS and equation congruence was much smaller and not significantly positive, $r = 0.19$ [$t(53) = 1.40$, $p = 0.17$]. This pattern of effects suggests that participants who held more positive attitudes towards statistics were more likely to examine and utilize the equation.

Discussion

Participants performed better on several measures of judgment accuracy, including proportion correct and $\overline{PS}$ when they had access to the statistical equation. However, they would have done far better still if they would have adhered to it consistently. In fact, participants in all cells would have showed improved performance if they had consistently used the best individual cue. Nevertheless, intuitive judgments matched the equation quite often, even under baseline (no access) conditions. This gives some credence to the general idea that human judgment is pretty good with naturalistic stimuli, or at least, not all bad.

We have seen that performance would have been improved if the aid had had more influence on judgment. Participants given access to the equation ensured there would be little chance of such influence by only infrequently inspecting the output from the aid. Why were they so reluctant to examine it? The results provide some support for two factors. First, a general attitude towards statistics seems to play a role, such that participants possessing a distaste for statistics look at the equation output less frequently. The second, and primary, factor under investigation is that participants believe that their intuitive judgment is better than it actually is. This claim was scrutinized in two ways. First, once the final test was completed, participants were asked to provide an aggregate or global estimate of their percent correct when their choices differed from that of the equation. Further, the instructions, scale, and procedure made it painfully clear that reporting that they were correct most of the time meant that they believed the equation was wrong most of the time. Even under these transparent, even leading, conditions participants on average reported that they were correct half or more of the time. These estimates were grossly overconfident.

The second approach to testing the overconfidence claim involved an (unsuccessful) attempt to reduce overconfidence experimentally. Specifically, some research suggests that experience that includes timely and reliable feedback leads to better calibration (e.g., Dougherty, 2001). Unfortunately for the experimental evaluation of our hypothesis, feedback did not reduce the overconfidence expressed during the task. Overconfidence was too incorrigible. Since feedback failed to reduce overconfidence, we have not yet determined whether or not a reduction in overconfidence leads to increased adherence to the statistical equation.

An alternative interpretation of the Experiment 1 results is that they are an artifact of lack of participant motivation. That is, it could be argued that the participants in Experiment 1 used the equation so infrequently because classroom credit provided insufficient motivation for them to care about their performance. A point against such an interpretation is that it is easier to completely ignore the juror profiles and not even try to formulate one's own judgment, but instead record the equation choice and confidence information verbatim. Indeed, Arkes et al. (1986) found that increasing participant motivation actually led to a decrease in equation usage. Nevertheless, we recruited paid participants in Experiment 2 to ensure that motivation was not a primary issue.

EXPERIMENT 2: PERFORMANCE MONITORING AND CALIBRATION FEEDBACK

In Experiment 1, outcome feedback failed to reduce overconfidence. This was unfortunate for the evaluation of our primary hypothesis. In particular, since overconfidence was not reduced, we cannot tell whether the reduction of overconfidence leads to greater reliance on a decision aid. Nevertheless, the experiment provided valuable information on the issue regarding the outcome feedback's lack of efficacy in reducing overconfidence. In Experiment 2, alternative methods for reducing overconfidence were attempted, as described presently.

In their extensive analysis of the relations between experience and confidence, Einhorn and Hogarth (1978) argued that confidence depends largely on the frequency of positive feedback, with the frequency of negative feedback playing a reduced role. Furthermore, the likelihood of observing positive feedback depends on factors such as judgmental ability and the base rate of successful prediction. In Experiment 1, the base rate of success was found to depend substantially on whether the individuals' judgments matched the equation; success was very unlikely when choices differed from the equation. Hence, encouraging participants to monitor their performance carefully, especially on those occasions where their choices differ from that of the equation, should lead to lower confidence.

An alternative, seemingly straightforward means of reducing overconfidence is to provide calibration feedback for a set of trials (e.g., Lichtenstein & Fischhoff, 1980). For example, after judging each of three to eleven sets of 200 items, participants in Lichtenstein and Fischhoff's (1980) studies were given extensive feedback and debriefing on their probability judgment accuracy, including measures of overconfidence and

graphical representations of their calibration. These experiments showed that a single session of 200 items followed by intensive feedback does lead to improved calibration, and that additional sessions yield little additional improvement.

In the current experiment, the performance monitoring and calibration feedback manipulations were varied independently. Outcome feedback was provided to all participants, because it was necessary for the performance monitoring manipulation. Providing it in the control condition was not expected to present any problem, since it was not found to improve calibration in Experiment 1. In addition, participants in Experiment 2 were financially compensated for their time to ensure reasonable levels of motivation.

Method

Participants

Study participants were 148 students and employees at Ohio State University who responded to a posted advertisement. They were each compensated $10 for their participation.

Design

The experiment was a 2 (performance monitoring; PM) $\times$ 2 (calibration feedback; CF) between-subjects design. All participants in Experiment 2 received outcome feedback in Block 1, and had access to the equation in both blocks. That is, the baseline conditions for Experiment 2 were identical to the feedback, equation access cell of Experiment 1. Block 2 was equivalent for participants in all conditions.

During Block 1, participants in the PM condition were asked to pause occasionally and reflect back on their performance, especially for those times when their choice differed from that of the equation. Every 40 trials (i.e., three times), they were specifically asked to respond ''yes'' or ''no'' to the following question: ''Please reflect back on your performance so far. When your predictions differed from the equation, were you correct at least 50% of the time?'' They also responded to the question immediately following the practice trials, so that they would know what to expect later on.

Participants in the CF condition received calibration feedback in graph form based on their Block 1 results, immediately prior to initiating Block 2. The calibration feedback consisted of percentage correct and average confidence in percentage form displayed in a bar graph that also showed the numeric values. It was explained that percentage correct and average confidence bars that were very close together implied good calibration, that confidence greater than percentage correct indicated overconfidence, and that confidence lower than percentage correct meant that the person was underconfident. Furthermore, the feedback was broken down and presented in three ways. First, results over all Block 1 trials were presented. Then, results from those trials where the respondent's choices matched the equation were presented. Finally, results for those trials where the respondent and equation disagreed were shown. While viewing each graph, respondents indicated whether their judgments showed good calibration (confidence within $\pm 3\%$ points of accuracy), overconfidence (confidence greater than accuracy by at least 3%), or underconfidence (confidence lower than accuracy by at least 3%). This was done to promote some level of understanding and absorption of the information presented in the graphs.

Procedure

The general procedure was essentially identical to that of Experiment 1.

Results

Confidence accuracy

Table 5 shows the means and standard deviations of the six measures for each condition, and Table 6 shows the inferential statistics. As shown, performance monitoring resulted in significantly higher slope, *r*, and a

Table 5. Means (SDs) for confidence and accuracy, Experiment 2

	Performance monitoring	
	No	Yes
No calibration feedback		
N	45	35
$\overline{PS}$	0.234 (0.042)	0.226 (0.029)
Confidence	0.74 (0.078)	0.75 (0.090)
Proportion correct	0.67 (0.067)	0.67 (0.054)
Bias	0.07 (0.098)	0.08 (0.0103)
Slope	0.03 (0.033)	0.05 (0.039)
Scatter	0.016 (0.007)	0.013 (0.006)
Pearson's r	0.13 (0.127)	0.20 (0.152)
Calibration feedback		
N	31	37
$\overline{PS}$	0.227 (0.040)	0.213 (0.020)
Confidence	0.73 (0.076)	0.74 (0.055)
Proportion correct	0.67 (0.075)	0.68 (0.039)
Bias	0.06 (0.101)	0.06 (0.055)
Slope	0.04 (0.042)	0.05 (0.037)
Scatter	0.014 (0.005)	0.015 (0.005)
Pearson's r	0.17 (0.158)	0.20 (0.123)

Table 6. F-statistics for confidence and accuracy, Experiment 2

	Factor		
Dependent variable	Calibration	Monitoring	Interaction
$\overline{PS}$	3.22	4.11*	0.32
Confidence	0.59	0.64	0.03
Proportion correct	0.31	0.99	0.83
Bias	0.99	0.00	0.53
Slope	1.45	6.05*	0.55
Scatter	0.01	1.08	3.49
Pearson's r	0.58	5.91*	1.26

*$p < 0.05$; **$p < 0.01$.

lower $\overline{PS}$. None of the other effects were statistically significant. In particular, and unfortunately for the evaluation of our hypothesis, neither performance monitoring, nor calibration feedback produced any reduction in overconfidence.

Equation inspection and congruence

Table 7 shows the means and standard deviations of the equation inspection and congruence measures. As can be seen, participants in the PM condition inspected the equation significantly more than those who did not monitor their performance, $F(1, 144) = 6.14$, $p = 0.014$. CF and the interaction were not statistically significant. Note that this effect for PM is not simply a ''manipulation check,'' as the PM group monitored their performance only in Block 1, and this effect is for Block 2. Participants' intuitions were fairly congruent with

Table 7. Means (SDs) for equation insepection and congruence, Experiment 2

	Performance monitoring	
	No	Yes
No calibration feedback		
Inspection	0.29 (0.356)	0.39 (0.389)
Match	0.81 (0.119)	0.83 (0.099)
Calibration feedback		
Inspection	0.26 (0.361)	0.48 (0.441)
Match	0.81 (0.127)	0.86 (0.074)

the equation, and corresponded well with the matching condition in Experiment 1. The slight increase in congruence for the PM condition was marginally significant, $F(1, 144) = 3.17$, $p = 0.077$.

Post-test questions

Retrospective aggregate confidence and overconfidence. The means and standard deviations of the aggregate confidence and bias values are presented in Table 8. As can be seen, for the choice-congruent estimate, aggregate confidence is increased and underconfidence is reduced significantly for the PM group, $F(1, 138) = 3.85$, $p = 0.052$, and $F(1, 138) = 4.36$, $p = 0.039$, respectively. Neither CF, nor the interaction was significant. No significant differences were found between conditions for the choice-incongruent estimate, though the effect for PM on overconfidence is marginal, $F(1, 139) = 2.81$, $p = 0.096$. As in Experiment 1, aggregate overconfidence is pervasive and sizeable across conditions for the choice-incongruent estimate, $t(142) = 10.29$, $p < 0.001$.

Table 8. Means (SDs) for aggregate confidence and bias (overconfidence). Experiment 2

	Performance monitoring	
	No	Yes
Choice congruent with equation		
No calibration feedback		
Agg. confidence	0.68 (0.181)	0.73 (0.112)
Agg. bias	−0.08 (0.182)	−0.03 (0.114)
Calibration feedback		
Agg. confidence	0.69 (0.164)	0.74 (0.121)
Agg. bias	−0.07 (0.167)	−0.02 (0.121)
Choice incongruent with equation		
No calibration feedback		
Agg. confidence	0.47 (0.174)	0.49 (0.155)
Agg. bias	0.17 (0.238)	0.26 (0.225)
Calibration feedback		
Agg. confidence	0.45 (0.204)	0.46 (0.186)
Agg. bias	0.17 (0.255)	0.21 (0.200)

Attitudes towards statistics. Across all conditions, scores on the ATS ranged from 45 to 95, with an overall mean of 71.41. There were no significant differences on ATS by condition. Since amount of inspection was found to differ by PM condition, correlations between ATS and inspection were examined separately by PM condition. For the non-PM conditions, the correlations between ATS score and equation inspection values were significantly positive, $r = 0.29$ [$t(74) = 2.58$, $p = 0.012$]. For the PM conditions, the correlation between ATS and equation inspection was essentially nil, $r = 0.02$. The overall correlation between ATS and equation congruence was positive and significant, $r = 0.21$ [$t(146) = 2.65$, $p = 0.009$]. These results suggest that PM influenced participants to examine the equation in a manner that was independent of their attitudes, thus negating the correlation. As in Experiment 1, those who held more positive attitudes towards statistics were generally more likely to utilize the equation.

Discussion

Participants who monitored their performance during the initial training session examined the equation more often in the test session and achieved better slope, *r*, and $\overline{PS}$ than those who were not prompted to monitor initial performance. However, performance monitoring did not reduce overconfidence during the task. Furthermore, the presentation of calibration feedback also failed to reduce overconfidence and, indeed, had no apparent effects on subsequent performance whatsoever. Hence, a clear experimental test of our primary hypothesis that overconfidence contributes to decision aid neglect went unrealized yet again. An adequate test requires that we are able to manipulate overconfidence, and doing so is not a trivial task.

The aggregate confidence estimates again revealed substantial overconfidence when choices differed from the equation, thus once more providing alternative support for the hypothesis that overconfidence contributes to decision aid neglect. Also, the finding that a relatively negative attitude towards statistics contributes to decision aid neglect was replicated as well. So, why did performance monitoring lead to increased reliance on the equation? And, why did calibration feedback fail to reduce overconfidence? These questions are addressed in turn.

With regards to the former question, we can only speculate at this point. First, one possibility is that the PM manipulation actually reduced overconfidence by the end of the Block 1 trials. The proposed reduction in overconfidence was then responsible for the increase in examination of the equation's output. Further, note that the equation's output included both choice and confidence values. To the extent that participants were then influenced by the equation's confidence values, their reported confidence would approach 75% (the equation's average confidence). Hence, we would not necessarily expect to find reduced confidence in Block 2, even if PM affected equation inspection by reducing confidence prior to Block 2. However, there is a point against the explanation that the PM effects were due to reduced overconfidence prior to Block 2, because confidence was not found to be lower in the PM than non-PM conditions at the tail end of the Block 1 trials. Also, there exists a possible alternative explanation. Recall that the performance monitoring task encouraged people to track their accuracy in the training session especially when their choice disagreed with the equation. Although not actually required, this strongly influenced people to examine the equation frequently during that initial training, thus giving them considerable exposure to the aid. And there is evidence that mere exposure to decision aids makes them generally more familiar and acceptable (Whitecotton, 1996). This possibility is potentially of considerable practical importance and will be addressed further in the general discussion.

With regards to the latter question, as noted, the calibration feedback manipulation's ineffectiveness surprised us, because Lichtenstein and Fischhoff (1980) found that calibration training did have some benefit. However, their training was more extensive than ours. Nevertheless, we thought that by highlighting those trials in which the equation and decision maker disagreed on the first block of trials, we would thereby be concentrating on those instances in which improvement in calibration would have a very large potential benefit. Rakow, Harvey, and Finer (2003) provided medical students with important base-rate information to help them decide which of a

sample of 36 medical school applicants were and were not admitted. The authors found that some medical students did not think that the base rate of acceptance of this sample was representative of the population, and other medical students simply did not take sufficient account of whatever base rate they thought was accurate. Similarly, in our study, some decision makers may have questioned whether the calibration feedback on Block 1 really pertained to Block 2 (''That must have been an usually tricky sample; this next group isn't likely to be that tough''), and others may have simply ignored whatever calibration feedback we provided (''I can't be that bad a judge''). Our third experiment was designed to make the calibration feedback more difficult to ignore, and to provide a clearer test of the focal hypothesis.

EXPERIMENT 3: EMBELLISHED CALIBRATION FEEDBACK

The goal of Experiment 3 was to provide a purer test of the hypothesis that overconfidence contributes to decision aid neglect. As discussed above, performance monitoring is suggestive, but the nature of the manipulation leaves it open to alternative explanation. In Experiment 3, we focus on calibration feedback provided between the training and test blocks, since effects due to it are unlikely to be attributable to anything aside from overconfidence reduction. And to be absolutely sure of the calibration feedback effect on performance when no equation is present, we included a manipulation check to show that calibration feedback reduces confidence, and has no impact on proportion correct. In the experiment proper, with the equation included at test, we expected to find the following results due to calibration feedback as compared with control:

(1) An increase in equation inspection and congruence.
(2) An increase in proportion correct.
(3) Same or decreased confidence.
(4) A net effect of reduced overconfidence.

Note that confidence may be the same because, as described above, increased inspection of the equation's output can lead to increased influence of its reported confidence values. However, in order to achieve this test, we first need to strengthen our calibration feedback manipulation. In particular, we wanted to be sure that the calibration feedback provided during the training phase of this experiment would be remembered and used during the test portion of the experiment. It may have been the case that the single question asked of participants in Experiment 2 resulted in only a temporary realization of how poor one's calibration was, but this realization had dissipated by the time testing during Block 2 occurred. In Experiment 3 we asked participants several questions about their calibration on the assumption that deeper processing of this information (Craik & Lockhart, 1972) would enhance participants' memory for it; thus its impact during the testing phase of the study (Block 2) would be greater than in Experiment 2. In addition, answering several questions about their calibration performance would require that participants would have to look at the calibration data multiple times.

Method

Participants

Study participants were 75 undergraduate students enrolled in an introductory psychology course at Ohio State University. Experimental participation was part of their course requirement.

Design

Enhanced calibration feedback (ECF) was the independent variable in Experiment 3, with participants either receiving it or not. In many respects the ECF was very much like the calibration feedback from

Experiment 2. The primary differences in Experiment 3 were, first, that participants responded to several questions from memory, after seeing each graph, rather than responding to a single question with the graph in view. In addition to the ''over-/under-confident'' question asked in Experiment 2, participants in the ECF condition entered their average confidence, percentage correct, the direction in which they should adjust their confidence to be well calibrated, and the amount of the adjustment needed. A button was included that allowed participants to go back and review the graph as needed. As before, the computer ensured that all entries were entered correctly prior to continuing with the experiment. Second, participants in the current experiment did not see graphs split on whether their choices matched the equation, since no equation had as yet been presented to them.

As a direct calibration feedback manipulation, the ECF condition should reduce confidence and not induce a change in proportion correct, as compared with a no-ECF control. In order to ensure that these were indeed the effects of the ECF, a small manipulation check was conducted that did not involve any introduction of the statistical equations. As anticipated, mean confidence was lower in the ECF condition (0.711) than in the control (0.746), $t(52) = 1.8$, $p = 0.039$ (one-tailed), and proportion correct was very similar in the two conditions (0.577 vs. 0.561), $t = 0.38$. Hence, the ECF appears to be an effective calibration feedback manipulation.[1]

Procedure

The procedure was very similar to those of Experiments 1 and 2. However, since no manipulations involving outcome feedback were incorporated, no outcome feedback was given during Block 1. Block 1 was also reduced to 60 trials because the experiment was running a bit long. Finally, the equation was not made available until Block 2, since its presence was unnecessary before then.

Results and discussion

Probability accuracy measures

Table 9 shows the means and standard deviations of $\overline{PS}$, confidence, proportion correct, bias, slope, scatter, and Pearson's r for each condition, along with the corresponding inferential statistics. As can be seen, ECF

Table 9. Means (SDs) for confidence and accuracy, Experiment 3

	Enhanced calibration feedback		
	No	Yes	t
N	37	38	—
$\overline{PS}$	0.230 (0.041)	0.216 (0.029)	1.75
Confidence	0.74 (0.038)	0.74 (0.051)	0.59
Proportion correct	0.64 (0.072)	0.68 (0.055)	2.32*
Bias	0.10 (0.076)	0.06 (0.064)	2.49*
Slope	0.05 (0.041)	0.05 (0.039)	0.04
Scatter	0.015 (0.007)	0.017 (0.009)	0.94
Pearson's r	0.21 (0.161)	0.19 (0.145)	0.39

*$p < 0.05$; **$p < 0.01$.

[1]As an aside, the proportion correct was lower here than in the no-equation condition of Experiment 1 because the test cases themselves happened to be harder by chance.

resulted in significantly higher proportion correct, and reduced overconfidence. None of the other effects were statistically significant.

Equation inspection and congruence

Participants in the control condition clicked the equation button on 48% of the trials on average, whereas those in the ECF condition clicked an average of 68% of the time, $t(73) = 2.32$, $p = 0.023$. Also, in the ECF condition participants' choices matched the equation on 87% of the trials, a significant gain over the 78% for those in the control condition, $t(73) = 3.15$, $p = 0.002$.

Post-test questions

Retrospective aggregate confidence and overconfidence. The means and standard deviations of the aggregate confidence and bias values for Experiment 3 are presented in Table 10. There are no significant differences between groups on these measures. As in the previous experiments, the levels of confidence and overconfidence associated with the choice-incongruent estimate are large and highly significant for both conditions, $t(33) = 8.37$, $p < 0.001$, and $t(35) = 7.76$, $p < 0.001$, respectively.

Attitudes towards statistics. Across conditions, scores on the ATS ranged from 54 to 93, with an overall mean of 73.35. There were no significant differences on ATS by condition. Although the ATS correlations were all positive, none were statistically significant in Experiment 3.

Summary

The ECF group was required to review and answer several questions about their calibration performance after completing their judgments in Block 1. A separate manipulation check indicated that this activity would reduce confidence but leave proportion correct unaffected when no equation was present. In congruence with our focal hypothesis, the ECF group viewed and matched the equation to a dramatically greater degree than did the control group. And their proportion of correct predictions was necessarily improved because of it. This, in turn, resulted in a diminution of the gap between confidence and accuracy for the ECF group. The complete set of results from these experiments are reviewed and elaborated upon in the following section.

Table 10. Means (SDs) for aggregate confidence and bias (overconfidence). Experiment 3

	Enhanced calibration feedback	
	No	Yes
Choice congruent with equation		
Agg. confidence	0.71 (0.178)	0.76 (0.151)
Agg. bias	−0.04 (0.183)	0.02 (0.159)
Choice incongruent with equation		
Agg. confidence	0.61 (0.216)	0.54 (0.216)
Agg. bias	0.36 (0.278)	0.34 (0.235)

GENERAL DISCUSSION

The results of the three experiments reported here provide considerable support for the proposals that statistical equations aid judgment, individuals overrate the prowess of their intuitive judgment abilities, and this overconfidence contributes to the neglect of these useful judgment aids. The latter was our focal hypothesis. Experiment 1 showed that access to a statistical equation resulted in improved judgment performance, but the benefits were compromised because those who had access to the equation's output often refused to even examine it. Furthermore, the participants exhibited marked overconfidence at the item level regardless of whether they had access to the equation. Also, those who had access to the equation reported at a global level that they performed far better when they went against the equation's recommendation than they actually did. This occurred despite global judgment instructions and procedures that made very salient the possibility that going against the equation could imply quite poor performance. The findings of low equation acceptability, improved performance with increased reliance, and overconfidence at both the item- and global-levels were replicated in Experiments 2 and 3. Experiment 3 further showed that reducing overconfidence via calibration feedback results in greater reliance on the equation, with a resulting improvement in performance. This summarizes the primary results pertaining to our core argument. Secondary results are mentioned below where pertinent.

Further appreciation for some of our results may be gleaned from the participants' open-ended responses provided at the end of the procedure. For example, one participant wrote, ''I went with intuition, and often used the equation for verification. If they did differ, however, I still went with my gut feeling. As the experiment went on and I got fatigued, I took the equation into account more.'' Brehmer (1972) and Slovic (1966) found that if cue information was in conflict (''He is very religious, but he drinks a lot''), then people gave inconsistent responses. Perhaps in our study when one's intuition and the equation were in conflict, the participant used the former sometimes (early in the experiment) but the latter at other times (later in the experiment), giving the same high confidence level each time. This would result in seriously overconfident judgments when one's intuition and the equation disagreed—one of the results that we found.

Another participant wrote, ''Statistical equation gave me more confidence if it was similar to my original guess. If it was different, I went with my gut instinct rather than using the equation. If I had absolutely no clue, I went with what the equation gave me.'' This statement is analogous to a result reported by Kahneman and Tversky (1973): people were much more likely to use base rates in making predictions if there was absolutely no individuating information which favored one alternative over another. In our study, there was abundant individuating information about each potential juror, so most decision makers had plenty of opportunity to use such information to override the equation if they were inclined to do so. If a gut instinct favored either option, that option was selected. Only if the gut was stymied was the statistical information used as a default. Of course, some participants, particularly those who had a negative attitude toward statistics, were less likely to use the equation under any circumstances: ''I chose to not use the statistical equation that much, especially during the second set of jurors, because I did not want it to influence my decisions too much.'' Lest college professors become too discouraged, we should highlight this response: ''The first thing I looked at was the statistical models answer. I was taught that a statistical model works better than most other forms of evaluation. So I weighted that more than the other options.'' The results of our study have clear implications for our primary issue of decision aid acceptability, as well as for the secondary issue of overconfidence. These issues are discussed in turn, beginning with the latter.

Three feedback-based methods for reducing overconfidence were attempted, all with pre-existing theoretical or empirical support. They included experience with regular outcome feedback, attention to negative outcome feedback, and two forms of aggregated calibration feedback. Of the three, only the stronger form of calibration feedback actually reduced overconfidence. The failure of regular outcome feedback to reduce overconfidence is of particular interest. Extant theories yield conflicting expectations on the matter (Dougherty, 2001; Einhorn & Hogarth, 1978; Sieck, 2003), though the prior data reviewed appears to favor

reduced overconfidence from outcome feedback (Dougherty, 2001; Fischhoff & Slovic, 1980). Future research is clearly needed with a particular view towards elucidating conditions under which outcome feedback will and will not reduce overconfidence, as well as comparing those results with the predictions of the competing theories. One plausible outcome is that the complete pattern of results will demand a thorough re-evaluation of all of the current focal theories.

A second method for reducing overconfidence involved instructions intended to direct attention to feedback in cases where negative outcome feedback was highly likely. This manipulation was ineffective. At first blush, this may appear to provide evidence that a lack of attention to negative feedback does not contribute to overconfidence. However, the null result provides only weak support at best for such a conclusion. The evidence is only damaging to the extent that participants were willing and able to follow the instruction, and that the instruction did not have any unanticipated consequences that mitigated its effectiveness (cf. Sanna, Schwarz, & Small, 2002). These issues were not addressed in the current study, because potential causes of overconfidence were not the primary topic under investigation. Nevertheless, they should be pursued in subsequent work. Finally, calibration feedback did reduce overconfidence, but only when combined with demands for extensive processing. In total, these results imply that overconfidence and poor judgment are not easily rectified (e.g., Gigerenzer, 1991; see also Griffin & Buehler, 1999).

The present experiments also point up the importance of overconfidence as a genuine phenomenon warranting serious investigation. Some researchers have argued that observed overconfidence can be attributed exclusively to statistical artifacts, and so is not a real psychological phenomenon (Juslin, Winman, & Olsson, 2000 but see Budescu, Wallsten, & Au, 1997). Research into this ''artifact'' issue has most directly proceeded by specifying psychological models that include explicit error terms, estimating the size of the error from behavioral data, and utilizing the error estimates to statistically adjust the data. The statistical adjustments tend to reduce overconfidence, with the remainder being taken to reflect ''real'' overconfidence. A largely unacknowledged problem with this approach is the underlying dubious assumption that the models and their associated constructs are veridical (Brenner, 2000). That is, the models are assigned a privileged status that is simply not warranted. The current research, along with a few other studies, can be viewed as taking a quite different approach to this problem (McGraw, Mellers, & Ritov, 2004; Renner & Renner, 2001). Specifically, these studies address whether overconfidence is a real phenomenon by exploring whether there are any real, practical implications of overconfidence. For example, Renner and Renner (2001) found that students were overconfident on introductory psychology quizzes, and that students in an experimental section who received calibration training achieved higher scores on subsequent tests than those in a control condition. McGraw et al. (2004) found that overconfidence was associated with negative affect, and that reducing overconfidence led to more positive emotional responses to a variety of tasks. Research along these lines suggests that overconfidence is indeed a very real psychological phenomenon (see also Koehler, Brenner, & Griffin, 2002).

An interesting feature of our results was that manipulations designed to reduce overconfidence were generally more effective in reducing retrospective overconfidence than they were in reducing item-by-item overconfidence. As Sniezek et al. (1990, p. 276) first pointed out, confidence in a single item may be determined in a different manner than confidence in a set of items. In particular, when staring at a single item, a person might invoke a compelling personal theory concerning the likelihood that a very religious non-drinker would favor physician-assisted suicide. Various forms of feedback, particularly aggregate-level feedback, might not diminish one's use of a personal theory given such a forceful cue. On the other hand, when retrospectively evaluating one's overall performance at the end of a long series of trials, a compelling but erroneous personal theory relating specific cues and physician-assisted suicide is no longer relevant in answering this aggregate question. The various feedback manipulations, therefore, do not have to overcome this serious obstacle to actuarial thinking, and a more reasonable estimate is forthcoming.

On the issue of decision aid acceptability, our results first reproduced the standard finding that statistical equations outperform human judges. This broadly replicable finding, along with the fact that linear statistical

models in particular possess very well understood theoretical properties, implies that they should generally be strongly preferred over unaided intuition. Nevertheless, as found here, judges rarely consult available equations and, even when they do consult, they routinely favor their gut feelings when the ''opinions'' differ. The current study yielded evidence in favor of three of the potential set of reasons for this state of affairs. First, as reviewed above, overconfidence was found to be an important factor. The results here corroborate existing work (Arkes et al., 1986; Whitecotton, 1996), and add strong additional support to the claim by showing experimentally that manipulating overconfidence produces changes in decision aid reliance. Taken together, the complete set of studies provides solid, unambiguous evidence that overconfidence is a critical aspect of the problem. However, until our understanding of the cognitive details underlying overconfidence improves considerably, there will be little practical value in recognizing this fact. Indeed, given the recalcitrance of overconfidence, it may be more fruitful from a practical stance to attempt to increase people's confidence in the equation, rather than attempting to decrease it in themselves. One way of accomplishing this might be to have participants write an exposition about why the equation is useful (Sieck & Yates, 1997). A second, novel, finding is that one's general attitude towards statistics may influence reliance, thus indicating that the usage decision is not merely based on confidence per se. Positive correlations were found between the ATS and reliance on the aid in Experiments 1 and 2. However, for reasons that we do not understand, the finding did not replicate in Experiment 3. Further work should be done to ascertain the reliability of this result. A more ambitious study might attempt ways of manipulating statistical attitudes to determine whether reliance can be leveraged in that way. Finally, in Experiment 2, we found that an experimental manipulation, which strongly encouraged participants to examine the aid frequently in the training session, led to increased examination and, to some extent, reliance on the aid in the final test. One possible mechanism underlying this result is that the added exposure to the equation's ''judgments'' during the training session increased participants' familiarity with that kind of output from the aid, thereby making it more palatable (cf. Zajonc, 1968).

Although the latter two of these ideas clearly require further testing, to the extent that they hold up, this set of factors suggests guidelines for slightly modifying statistical education programs towards rectifying the current state of affairs. First, a short version of a judgment task along the lines of the enhanced calibration condition could be incorporated into the curriculum to illuminate individuals' overconfidence in their intuitive judgment, and the effectiveness of statistical models. Second, teachers ought to adopt the development of positive attitudes towards statistics as an explicit course objective. Finally, in addition to having students study the mechanics of statistical models, there may be a real additional benefit in having them practice generating model predictions, and using those predictions to judge individual cases. Adjustments such as these to current educational practice ought to be tried and evaluated so that future professionals are more mentally prepared to adopt the best judgment practices available.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grants SBR-0196200 and SES-0326468.

REFERENCES

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: what consumers know and what they think they know. *Journal of Consumer Research, 27*, 123–156.

Alvarado, A. (1986). A practical score for the early diagnosis of acute appendicitis. *Annals of Emergency Medicine, 15*(5), 79–86.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods for reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39*, 133–144.

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes, 37*, 93–110.

Arkes, H. R., Dawson, N. V., Speroff, T., Harrell, F. E., Alzola, C., Phillips, R., Desbiens, N., Oye, R. K., Knaus, W., & Connors, A. F. (1995). The covariance decomposition of the probability score and its use in evaluating prognostic estimates. *Medical Decision Making, 15*, 120–131.

Ashton, R. (1991). Pressure and performance in accounting decision settings: paradoxical effects of incentives, feedback, and justification. *Journal of Accounting Research, 28*, 148–186.

Brehmer, B. (1972). Cue utilization and cue consistency in multiple-cue probability learning. *Organizational Behavior and Human Performance, 8*, 286–296.

Brehmer, B. (1980). In one word: not from experience. *Acta Psychologica, 45*, 223–241.

Brenner, L. (2000). Should observed overconfidence be dismissed as a statistical artifact? Critique of Erev, Wallsten, and Budescu (1994). *Psychological Review, 107*(4), 943–946.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997). On the importance of random error in the study of probability judgment. Part II: applying the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making, 10*(3), 173–188.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: a theoretical framework and implications for science instruction. *Review of Educational Research, 63*, 1–49.

Corey, G. A., & Merenstein, J. H. (1987). Applying the acute ischemic heart disease predictive instrument. *The Journal of Family Practice, 25*, 127–133.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671–684.

Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist, 34*, 571–582.

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General, 130*, 579–599.

Dougherty, 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*(5), 395–416.

Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting, 1*, 155–172.

Fischhoff, B., & Slovic, P. (1980). A little learning . . . : confidence in multicue judgment tasks. In R. Nickerson (Ed.), *Attention and Performance* (Vol. VIII, pp. 779–800). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond ''Heuristics and Biases''. In W. Stroebe, & M. Hewstone (Eds.), *European Review of Social Psychology* (Vol. 2). Chichester: Wiley.

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

Gigerenzer, G., Todd, P. M., & ABC Group. (1999). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Graham, I. D., Stiell, I. G., Laupacis, A., McAuley, L., Howell, M., Clancy, M., Durieux, P., Simon, N., Emparanza, J. I., Aginaga, J. R., O'Connor, A., & Wells, G. (2001). Awareness and use of the Ottawa Ankle and Knee Rules in 5 countries: can publication alone be enough to change practice? *Annals of Emergency Medicine, 37*(3), 259–266.

Griffin, D., & Buehler, R. (1999). Frequency, probability, and prediction: easy solutions to cognitive illusions? *Cognitive Psychology, 38*, 48–78.

Hammond, K. R., Summers, D. A., & Deane, D. H. (1973). Negative effects of outcome-feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance, 9*, 30–34.

Hoffman, P. J., Slovic, P., & Rorer, L. G. (1968). An analysis-of-variance model for the assessment of configural cue utilization in clincal judgment. *Psychological Bulletin, 69*, 338–349.

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*(2), 384–396.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Keren, G. (1987). Facing uncertainty in the game of bridge: a calibration study. *Organizational Behavior and Human Decision Processes, 39*(39), 98–114.

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. 1–18). Cambridge, MA: Cambridge University Press.

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

Levin, M. (1975). *A Cognitive Theory of Learning: Research on Hypothesis Testing*. Hillsdale, NJ: Erlbaum.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*, 149–171.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: theories and models 1980–94. In G. Wright, & P. Ayton (Eds.), *Subjective Probability* (pp. 453–482). Chichester: Wiley.
McGraw, A. P., Mellers, B., & Ritov, I. (2004). The affective costs of overconfidence. *Journal of Behavioral Decision Making, 17*, 281–295.
McNiel, D. E., & Binder, R. L. (1994). Screening for risk of inpatient violence: validation of an actuarial tool. *Law and Human Behavior, 18*, 579–586.
Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*, 370–375.
Nelson, T. O. (1984). A comparison of current measures of feeling-of-knowing accuracy. *Psychological Bulletin, 95*, 109–133.
Peterson, D. K., & Pitz, G. F. (1986). Effect of input from a mechanical model on clinical judgment. *Journal of Applied Psychology, 71*(1), 163–167.
Price, P. C. (1998). Effects of a relative-frequency elicitation question on likelihood judgment accuracy: the case of external correspondence. *Organizational Behavior and Human Decision Processes, 76*(3), 277–297.
Rakow, T., Harvey, N., & Finer, S. (2003). Improving calibration without training: the role of task information. *Applied Cognitive Psychology, 17*, 419–441.
Renner, C. H., & Renner, M. J. (2001). But I thought I knew that: using confidence estimation as a debiasing technique to improve classroom performance. *Applied Cognitive Psychology, 15*, 23–32.
Sanna, L. J., Schwarz, N., & Small, E. M. (2002). Accessibility experiences and the hindsight bias: I knew it all along versus it could never have happened. *Memory & Cognition, 30*(8), 1288–1296.
Schultz, K. S., & Koshino, H. (1998). Evidence of reliability and validity for Wise's Attitudes Toward Statistics scale. *Psychological Reports, 82*, 27–31.
Sieck, W. R. (2003). Effects of choice and relative frequency elicitation on overconfidence: further tests of an exemplar-retrieval model. *Journal of Behavioral Decision Making, 16*, 127–145.
Sieck, W. R., & Yates, J. F. (1997). Exposition effects on decision making: choice and confidence in choice. *Organizational Behavior and Human Decision Processes, 70*, 207–219.
Sieck, W. R., & Yates, J. F. (2001). Overconfidence effects in category learning: a comparison of connectionist and exemplar memory models. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27*(4), 1003–1021.
Slovic, P. (1966). Cue-consistency and cue-utilization in judgment. *American Journal of Psychology, 79*, 427–434.
Slovic, P. (1995). The construction of preference. *American Psychologist, 50*, 364–371.
Sniezek, J. A., Paese, P. W., & Switzer, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes, 46*, 264–282.
Whitecotton, S. M. (1996). The effects of experience and confidence on decision aid reliance: a causal model. *Behavioral Research in Accounting, 8*, 194–216.
Wilkie, M. E., & Pollock, A. C. (1996). An application of probability judgment accuracy measures to currency forecasting. *International Journal of Forecasting, 12*, 25–40.
Wise, S. L. (1985). The development and validation of a scale measuring attitudes toward statistics. *Educational and Psychological Measurement, 45*, 401–405.
Yates, J. F. (1982). External correspondence: decompositions of the mean probability score. *Organizational Behavior and Human Decision Processes, 30*, 132–156.
Yates, J. F. (1990). *Judgment and Decision Making*. Englewood Cliffs, NJ: Prentice Hall.
Yates, J. F. (1994). Subjective probability accuracy analysis. In G. W. P. Ayton (Ed.), *Subjective Probability* (pp. 381–410). New York: Wiley.
Yates, J. F., & Estin, P. (1996, November). *Training Good Judgment.* Paper presented at the Annual Meeting of the Society for Judgment and Decision Making, Chicago.
Yates, J. F., Lee, J.-W., & Shinotsuka, H. (1992, November). *Cross-national Variation in Probability Judgment.* Paper presented at the Annual Meeting of the Psychonomic Society, St. Louis.
Yates, J. F., Lee, J.-W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: beyond general knowledge overconfidence? *Organizational Behavior and Human Decision Processes, 74*(2), 89–117.
Yates, J. F., Lee, J.-W., Sieck, W. R., Choi, I., & Price, P. C. (2002). Probability judgment across cultures. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 271–291). Cambridge: Cambridge University Press.
Yates, J. F., Veinott, E. S., & Patalano, A. L. (2003). Hard decisions, bad decisions: on decision quality and decision aiding. In S. L. Schneider, & J. Shanteau (Eds.), *Emerging Perspectives on Judgment and Decision Research* (pp. 13–63). New York: Cambridge University Press.
Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monographs, 9*, 1–27.

Authors' biographies:

Winston R. Sieck received his PhD in cognitive psychology from the University of Michigan in 2000, served in quantitative psychology at Ohio State University, and is currently a naturalistic researcher at Klein Associates. He is interested in investigating practical and theoretical issues in human decision making from a variety of perspectives.

Hal R. Arkes is a professor of psychology and Interim Director of the Center for Health Outcomes, Policy and Evaluation Studies at Ohio State University. He received his PhD in 1971 from the University of Michigan. His research interests are in the areas of medical and economic decision making.

Authors' addresses:

Winston R. Sieck, 1750 Commerce Center Blvd. North, Fairborn, OH 45324-3987, USA.

Hal R. Arkes, Department of Psychology, Ohio State University, 1827 Neil Avenue, Columbus, OH 43210-1222, USA.

