Journal of Experimental Psychology:
Learning, Memory, and Cognition
2004, Vol. 30, No. 5, 947–959

0278-7393/04/$12.00 DOI: 10.1037/0278-7393.30.5.947

Explaining Purportedly Irrational Behavior by Modeling Skepticism in Task Parameters: An Example Examining Confidence in Forced-Choice Tasks

Craig R. M. McKenzie and John T. Wixted
University of California, San Diego

David C. Noelle
Vanderbilt University

Many purported demonstrations of irrational behavior rely on the assumption that participants believe key task parameters that are merely asserted by experimenters. For example, previous researchers have found that participants who first reported confidence in items presented in a yes–no format did not change confidence to the degree prescribed by the normative model when those same items were later presented in a forced-choice format. A crucial assumption, however, was that participants fully believed the assertion that the forced-choice items were mutually exclusive and exhaustive. In this article, the authors derive and test a new normative model in which it is not assumed that participants fully believe the assertion. Two visual identification experiments show that the new normative model provides a compelling account of participants' confidence reports.

In essence, all information is imperfectly reliable, and information provided to participants by experimenters is no exception. Indeed, experimenter-provided information might be less trustworthy than most. Participants are often deceived in psychology experiments, and they are aware of this. Not only do students routinely learn about classic studies in which deception has been used (e.g., Asch, 1955; Milgram, 1963), but the American Psychological Association (APA) requires that participants be fully debriefed whenever they are deceived (APA, 1992). First- or secondhand knowledge about deception in psychology experiments is widespread.

Concerned about the use of deception in psychology experiments, Hertwig and Ortmann (2001; see also Ortmann & Hertwig, 1997) noted a variety of ways in which suspicious participants behave differently than trusting ones. Not mentioned by these authors, however, is what we consider to be an especially troubling aspect of the implications of participant skepticism. In many tasks, participants' behavior is compared to a normative standard, and differences between behavior and the normative response are routinely interpreted as errors. Often, however, a critical assumption in attributing errors to participants is that they believe the assumptions underlying the purported normative response (e.g., that the options are mutually exclusive and exhaustive, or that the observations are randomly sampled). If participants do not fully believe key task parameters, which are often merely asserted by experimenters, then calling their responses "errors" would be misleading.

Indeed, findings by other authors suggest that increasing the believability of important task parameters can change participants' behavior in the direction of the normative response. Although we investigate a different task, some previous findings regarding participants' use of base rates provide an instructive example. On the basis of results from their well-known lawyer–engineer problem, Kahneman and Tversky (1973) argued that participants were committing a normative error by severely underweighting base rates when reporting subjective probabilities. However, Gigerenzer, Hell, and Blank (1988) disputed that conclusion, because a key assumption underlying the normative model—that the observations on which the subjective probabilities were based were randomly sampled—was merely asserted by the experimenters. Gigerenzer et al. made random sampling seem more realistic to participants by having them draw the observations themselves from urns. This resulted in subjective probabilities that were much more similar to the normative responses.

The fact that increasing the realism (and hence believability) of a key task parameter can lead to more normative responses suggests that participant skepticism should be taken seriously as a reason why purported errors occur. However, without knowing participants' *degree of belief* in the parameter of interest and how they ought to respond given that degree of belief, important questions will inevitably remain unanswered. For example, even the responses of Gigerenzer et al.'s (1988) participants who sampled the observations themselves deviated from the purportedly rational model. Was this because they were making errors (albeit relatively small ones), or were they responding optimally given residual doubts about the legitimacy of the random sampling procedure

Craig R. M. McKenzie and John T. Wixted, Department of Psychology, University of California, San Diego; David C. Noelle, Department of Electrical Engineering and Computer Science, Vanderbilt University.

This research was supported by National Science Foundation Grants SBR-9515030, SES-0079615, and SES-0242049, and some of the results were presented at the 2002 annual meeting of the Psychonomic Society, Kansas City, Missouri, November 2002. We thank Michael Liersch, Shlomi Sher, and Trish Van Zandt for their helpful comments.

Correspondence concerning this article should be addressed to Craig R. M. McKenzie, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, MC 0109, La Jolla, CA 92093-0109. E-mail: cmckenzie@ucsd.edu

(which was, in fact, illusory)? Perhaps it is not reasonable to assume that participants fully believe key task parameters, even when experimenters go to great lengths to increase believability. (Participants still know, after all, that they are participating in a psychology experiment.) Furthermore, if Kahneman and Tversky's (1973) participants had little or no faith in the random sampling claim, might even they have been responding optimally?

In this article, we provide a specific example of a general approach to dealing with participant skepticism that can address the sorts of questions raised above. In general, we propose that researchers derive and test *normative models* that do not assume full belief in important task parameters. Although the approach is quite general in theory, it must be implemented on a task-by-task basis in practice. The topic we examine is confidence in forced-choice tasks, which has purportedly been shown to be suboptimal. More specifically, it has been shown that, when presented with a pair of forced-choice items and asked for confidence that a particular (focal) item is the true one, participants are insufficiently influenced by their confidence that the nonfocal item is the true one (as determined by their earlier responses in a yes–no format in which the same items were presented individually; McKenzie, Wixted, Noelle, & Gyurjyan, 2001). However, a crucial assumption underlying the claim of suboptimality is that participants believe that the forced-choice options are mutually exclusive and exhaustive (i.e., one item is true and the other false). Below, we discuss how, in the current studies, we derived and tested a normative model that does not assume full belief in this parameter, allowing us to check whether participants are responding rationally, given their less-than-full belief that the forced-choice options are mutually exclusive and exhaustive.

The current approach treats participant skepticism as an inevitable and *tractable* variable in psychology experiments and has the potential for leading to a better understanding of behavior, especially when claims about errors are made in tasks in which a crucial assumption is that participants believe certain task parameters. To our knowledge, we are the first to adopt such an approach.

The rest of the article is organized as follows. In the next section, we describe recent research examining change in confidence between yes–no and forced-choice tasks and how participants' responses deviate from the purported normative model. In the subsequent section, we introduce a new normative model in which it is not assumed that participants believe that the forced-choice alternatives are mutually exclusive and exhaustive. We then report results from two experiments that test the traditional normative model, our new normative model, and a descriptive model using a visual identification task. In the General Discussion section, we discuss potential implications for other areas of research in which researchers assume that participants believe key task parameters.

Relation Between Confidence in Yes–No Tasks and Confidence in Forced-Choice Tasks

Yes–no and forced-choice tasks are widely used in psychology. In perception experiments, for example, participants are sometimes asked to decide whether a particular stimulus appeared (a yes–no task) or to decide which one of two or more stimuli appeared (a forced-choice task). Similarly, in categorization experiments, participants might be asked to decide whether an item belongs to a particular category (yes or no) or to decide which of two categories the item belongs to (forced choice). In recognition memory experiments, the task is to decide whether an individual item appeared on an earlier study list or to decide which member of a pair of items appeared on the list. Finally, in judgment and decision making tasks, participants might report whether an individual statement is true or select the one true answer from among multiple alternatives. In both yes–no and forced-choice tasks, participants sometimes report their confidence that they made the correct response.

Despite the ubiquity of the two tasks, the issue of how confidence between them is related has only recently been examined empirically (McKenzie et al., 2001). Consider being presented sequentially with two general-knowledge statements and reporting confidence that each is true: (A) The population of the United States is greater than 265 million. (B) Sophocles was born before Socrates. Assume that you are 80% confident that A is true and 40% confident that B is true. You are then presented with A and B simultaneously and told that exactly one of the statements is true—that is, the task has changed from yes–no for each of A and B to forced choice involving both A and B. Now how confident would you be that A is true? That B is true? How confident *should* you be?

A Normative Model and a Descriptive Model

To illustrate the normative forced-choice response, continue assuming that yes–no confidence in A and B is 80% and 40%, respectively. These probabilities (and their complements) are represented by the marginals in Figure 1. When A and B are subsequently paired, one of four possibilities can occur: Both A and B are true (A&B), A is true and B false (A&~B), A is false and B true (~A&B), and both A and B are false (~A&~B). We assume throughout this article that confidence in A and confidence in B are statistically independent at the yes–no stage, allowing us to simply multiply the marginals to calculate the probability of each of the four possible joint outcomes, shown in the respective cells.

The typical forced-choice task involving A and B is one in which the A&B and the ~A&~B outcomes are impossible. Either A is true and B is false, or A is false and B is true. In frequentist terms, because 56 of 100 outcomes will fall under the A&~B and ~A&B categories, and 48 belong to the former, the probability that A is true, assuming that A and B are mutually exclusive and exhaustive, is 48/56, or about .86. Similarly, the probability that B is true is 8/56, or .14. In the example, then, forced-choice confidence in A should increase from 80% to 86%, and confidence in B should decrease from 40% to 14%. The following is the normative model, given the above assumptions (McKenzie et al., 2001; see also Ferrel & McGoey, 1980; Luce, 1963):

$$c(A,B) = c(A)[1 - c(B)]/\{c(A)[1 - c(B)] + c(B)[1 - c(A)]\} \quad (1)$$

where $c(A,B)$ corresponds to confidence that A is true given that A and B are mutually exclusive and exhaustive (i.e., at the forced-choice stage), and where $c(A)$ and $c(B)$ are confidence in A and B

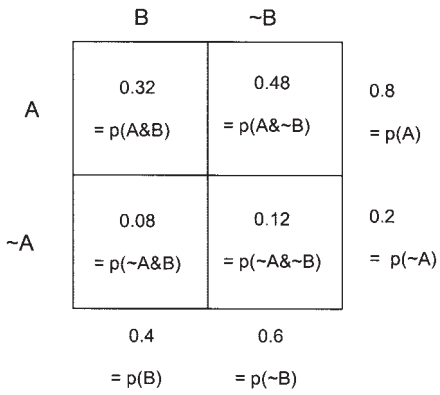

Figure 1. An illustration of the normative model in 2 × 2 form using a general-knowledge task (as described in the text). The marginal probabilities correspond to confidence in statements A and B (and their complements) in a yes–no task. When the statements are paired, one of four outcomes can occur: (a) Both A and B can be true (upper left cell), (b) A can be true and B false (upper right), (c) B can be true and A false (lower left), or (d) both A and B can be false (lower right). The resulting (joint) probabilities in each cell assume that A and B are statistically independent. In a forced-choice task involving A and B, the A&B and the ~A&~B outcomes are not possible: Either A or B is true, but not both. For the forced-choice task, then, confidence in A should increase from 80% to 0.48/(0.48 + 0.08) = 86%. Similarly, confidence in B should decrease from 40% to 0.08/(0.08 + 0.48) = 14%.

when they are independent (i.e., at the yes–no stage). (McKenzie et al. also present the normative model without the assumption that confidence levels in the yes–no items are independent.) To calculate c(*B*,*A*), simply switch c(*A*) and c(*B*) in Equation 1.

The above normative model describes how participants should change confidence between yes–no and forced-choice tasks. McKenzie et al. (2001) also tested three descriptive models and, overall, the best-performing one was a more general form of the normative model, which they referred to as the multiplicative model. The model is much like the normative model, except that one free parameter is added to the equation:

$$c(A,B) = c(A)[1 - c(B)]^w / \{c(A)[1 - c(B)]^w + c(B)^w[1 - c(A)]\}. \quad (2)$$

The parameter, *w*, is associated with each term involving c(*B*) and determines the extent to which, holding c(*A*) and c(*B*) constant, c(*B*) affects c(*A*,*B*). The expected range of *w* is 0–1. When $w = 1$, Equations 1 and 2 are identical and, as *w* decreases, c(*B*) has a smaller effect on c(*A*,*B*). When $w = 0$, c(*B*) has no effect on c(*A*,*B*), which will then equal c(*A*). Note that participants' responses are normative only when $w = 1$ in this model. Considerable evidence indicates that nonfocal alternatives are often underweighted, so there is good reason to believe that *w* might be less than 1 (e.g., Evans, 1989; Fischhoff & Beyth-Marom, 1983; Klayman & Ha, 1987; McKenzie, 1994, 1998, 1999). If $w < 1$, then confidence in A and B in the forced-choice case will not sum to 100%, except when c(*A*) and c(*B*), confidence at the yes–no stage, happen to sum to 100%. The normative model, in contrast, always requires that forced-choice confidence sum to 100%.

Previous Empirical Findings

McKenzie et al. (2001) reported two experiments, but we just describe one, because both led to similar findings. Their Experiment 2 presented participants with general-knowledge statements. Confidence was reported in the truth of each individual statement and then was reported again when the statements had been put into pairs such that one statement was true and one was false. Half of the participants were instructed to have their forced-choice confidence reports sum to 100, and half were not. A general finding was that the normative model (Equation 1) did not accurately predict participants' responses (see also Bender, 1998), but that the multiplicative model fit well. Furthermore, the value of the multiplicative model's *w* parameter that provided the best fit was always less than 1, indicating that confidence in nonfocal alternatives was underweighted. McKenzie et al. noted that one reason for the suboptimal impact is that participants might not have fully believed the assertion that the forced-choice items were mutually exclusive and exhaustive, but they did not pursue the issue. In the next section, we describe a new normative model that can help determine the viability of this account.

Trust Model

The normative model (Equation 1) assumes that the participant fully believes, at the forced-choice stage, that A and B are mutually exclusive and exhaustive. However, participants could believe to some extent that A and B might both be true and/or that they might both be false. As mentioned, participants often distrust what experimenters tell them, usually with good reason (e.g., Hertwig & Ortmann, 2001; Ortmann & Hertwig, 1997). In fact, participants are sometimes deceived about forced-choice items being mutually exclusive and exhaustive. In the memory literature, for example, Glanzer and Bowles (1976) told participants that they would be presented with a target and a lure, but null test trials, which involved a choice either between two targets or between two lures, were also included without warning. This procedure has since been used quite frequently (e.g., Wixted, 1992).

What is the optimal level of confidence to report in an item at the forced-choice stage if the reliability of the experimenter's assertion that the two alternatives are mutually exclusive and exhaustive is considered to be less than 100%? The answer depends on the degree to which the experimenter's assertion is believed. How can this be determined? A reasonable starting point is to ask what the probability is that A and B are mutually exclusive and exhaustive prior to a consideration of what the experimenter has to say about it. Whatever that prior probability turns out to be, it would then be updated, depending on how reliable the participant considers the experimenter to be.

The prior probability that the forced-choice alternatives are mutually exclusive and exhaustive depends on c(*A*) and c(*B*), the level of confidence in the two alternatives when they are considered independently (i.e., at the yes–no stage). If, for example, the participant is highly confident that A is true and highly confident that B is true (or that both are false) when they are considered independently, then it should seem unlikely that one is true and the other false when A and B are presented as a pair. In contrast, if a participant has high confidence in A and low confidence in B (or

vice versa), then it should seem likely that one is true and the other false when they are presented as a pair. More precisely, the prior probability, q, that two alternatives are mutually exclusive and exhaustive is given by the following:

$$q = c(A)[1 - c(B)] + c(B)[1 - c(A)]. \quad (3)$$

If, as in our previous example, $c(A) = 80\%$ and $c(B) = 40\%$, then the probability that A is true and B false (A&~B) is $0.8 \times 0.6 = 0.48$, and the probability that B is true and A false (~A&B) is $0.2 \times 0.4 = 0.08$. In this case, q equals $0.48 + 0.08 = 0.56$. Thus, prior to taking into account the information supplied by the experimenter, a normative participant would assume that the probability is .56 that one of the two alternatives is true and the other false.

The prior probability is then updated on the basis of the information supplied by the experimenter (who claims that the alternatives are mutually exclusive and exhaustive) in accordance with Bayes's theorem. We assume that the degree to which the prior probability changes in response to that information is a function of the experimenter's perceived reliability. The more reliable the experimenter is perceived to be, the more the prior probability is adjusted upward.

The following is Bayes's theorem using present notation:

$$s = qr/[qr + (1 - q)(1 - r)], \quad (4)$$

where q represents the prior probability that the alternatives are mutually exclusive and exhaustive, or p(MEE), s represents the posterior probability (i.e., the probability that the two alternatives are mutually exclusive and exhaustive after taking into account that the experimenter has said they are), or p(MEE|"MEE"), and r represents experimenter reliability, or p("MEE"|MEE). More specifically, r represents the probability that the experimenter would claim that the two alternatives are mutually exclusive and exhaustive given that they really are, and $1 - r$ is the probability that the experimenter would claim that the two alternatives are mutually exclusive and exhaustive given that they are not.[1]

For a participant convinced that the information provided by the experimenter is 100% reliable ($r = 1$), Equation 4 shows that the posterior probability, s, is equal to 1 and that normative performance is prescribed by Equation 1. However, for a participant who believes that the information provided by the experimenter is less than 100% reliable, normative performance is given by the following, closely related, equation (derived in Appendix A):

$$c(A,B) = w(c(A)[1 - c(B)]/\{c(A)[1 - c(B)] + c(B)[1 - c(A)]\}) + (1 - w)c(A), \quad (5)$$

where $w = (s - q)/(1 - q)$. In words, Equation 5 states that confidence in A in the forced-choice situation is a weighted average of the normative model (Equation 1) and confidence in A at the yes–no stage. For a participant who believes that the experimenter is 100% reliable, $s = 1$ and, therefore, $w = 1$. In that case, Equation 5 reduces to Equation 1. For a participant who believes that the experimenter is only 50% reliable (i.e., $r = .5$ such that the experimenter's claim adds no new information), $s = q$, according to Equation 4, and, therefore, $w = 0$. In that case, Equation 5 reduces to $c(A)$, which is to say that the participant's confidence in A in the forced-choice situation remains what it was in the yes–no situation (as it should, if the experimenter's assertion is considered to be completely uninformative). Finally, for a participant who believes that the experimenter is of intermediate reliability (i.e., $.5 < r < 1$), confidence in A will fall somewhere between $c(A)$ and what Equation 1 says it should be. That is, the change in confidence in A should be somewhat less than is prescribed by Equation 1. Equation 5 is what we call the *trust model* because its prediction depends on the degree to which the participant believes that the forced-choice items are mutually exclusive and exhaustive. When fitting the model to participants' data, we assume that r is determined prior to participants' seeing any forced-choice items and does not vary with the particular pair of items under consideration.

The trust model is interesting for two reasons. First, it is normative under the assumption that the judge does not fully believe that A and B are mutually exclusive and exhaustive (although, for the sake of simplicity, we refer to Equation 1 as the normative model unless noted otherwise). Second, the model shows that it can be normatively appropriate for $c(B)$ to have less impact on $c(A,B)$ than is prescribed by Equation 1.

As written, the trust model is conceptually appealing, but it is written in terms of q and s (the prior and posterior probabilities that A and B are mutually exclusive and exhaustive). To estimate the degree to which a participant considers the information provided by the experimenter to be reliable, an equation that contains r as a free parameter is needed. This is accomplished by substituting the right side of Equation 4 for s in Equation 5 and the right side of Equation 3 for q in Equation 5. The result (see Appendix B) is

$$c(A,B) = \frac{rc(A)[1 - c(B)] + (1 - r)c(A)c(B)}{r\{c(A)[1 - c(B)] + c(B)[1 - c(A)]\} + (1 - r)\{c(A)c(B) + [1 - c(A)][1 - c(B)]\}}. \quad (6)$$

The right-hand side of Equation 6 differs from the normative model (Equation 1) in part because it includes the terms $c(A)c(B)$ and $[1 - c(A)][1 - c(B)]$. The former term represents the probability that A and B are both true, and the latter represents the probability that A and B are both false. In Equation 6, we assign a weight to these two terms that reflects the extent to which the participant believes that those two possibilities still exist in spite of what the experimenter has said. When $r = .5$, the terms are included completely, and $c(A,B) = c(A)$. When $r = 1$, the terms are excluded, and it is easy to see that Equation 6 reduces to Equation 1.

[1] For the sake of exposition, we are assuming that the two likelihoods, p("MEE"|MEE) and p("MEE"|~MEE), sum to 1, but we do not make this assumption in the formulation of our models. In other words, in the simplified text, we are assuming that the experimenter is unbiased when claiming that the forced-choice items are or are not mutually exclusive and exhaustive. To see that we are not limited by this assumption, however, note that experimenter reliability can be represented by the likelihood ratio (LR) in Bayes's theorem, p("MEE"|MEE)/p("MEE"|~MEE). LR in the current context is expected to be at least 1. We assume that the more reliable the experimenter is perceived to be, the greater LR is. We can rescale this value to range between 0.5 and 1: LR/(LR + 1). The resulting value is r. Note that $r/(1 - r) = LR$. Thus, Equation 4 holds even for cases in which the two likelihoods do not sum to 1, or in which the experimenter has a bias for the saying that the items are (or are not) mutually exclusive and exhaustive.

Because Equations 5 and 6 are algebraically equivalent (see Appendix B), the trust model can be represented by either one. In Equation 6, c(*A*) and c(*B*) for a given pair define everything in the equation needed to compute the predicted value of c(*A*,*B*) except for *r*. That value must be estimated from a participant's data. In the experiments described in this article, the data from an individual participant will involve many c(*A*), c(*B*), c(*A*,*B*) combinations (i.e., many cases in which the participant has supplied confidence in A and B independently and in a subsequent forced-choice arrangement). Equation 6 can be fit to such data by adjusting the value of *r* to yield predicted values of c(*A*,*B*) that are as close as possible to the observed values. This is a sensible thing to do because, theoretically, *r* is constant across all of the pairs under consideration. To the extent that Equation 6 fits the data better than competing models and that *r* varies in the expected way when degree of belief changes with respect to the forced-choice items being mutually exclusive and exhaustive, the more compelling the trust model will be.[2]

The Relationship Between the Multiplicative and Trust Models

The multiplicative and trust models are conceptually different. The former is a descriptive account that implies that confidence in nonfocal alternatives has less than optimal impact (relative to Equation 1) for whatever reason. The latter implies that the suboptimal impact is specifically due to failure to believe fully that the items are mutually exclusive and exhaustive. Despite the conceptual differences, the two models are nearly identical in their quantitative behavior. In particular, as our experiments show, the two models result in virtually identical goodness-of-fit statistics.

Because of the quantitative similarity between the multiplicative and trust models, the findings reported by McKenzie et al. (2001) in support of the multiplicative model can also be taken to support the trust model. That is, in terms of goodness of fit, the trust model outperforms the normative model. Although they provide nearly identical quantitative fits, the advantage of the trust model is that it provides a specific reason why *w* in the multiplicative model was less than 1. The multiplicative model is silent about that. The question of interest is whether the explanation offered by the trust model is not only specific but correct. If the trust model is correct, then skepticism regarding the experimenter's claim that the forced-choice items are mutually exclusive and exhaustive should influence *r* in predictable ways.

Experiment 1

Our primary purpose in Experiment 1 was to examine the trust model's conceptual accuracy by asking participants, after the experiment was completed, about their belief that the targets during the forced-choice stage were mutually exclusive and exhaustive. Support for the model's conceptual accuracy would be found to the extent that *r* is higher for those who did versus those who did not believe that the targets were mutually exclusive and exhaustive.

In addition, Experiment 1 allowed for testing whether the performance of the models using general-knowledge statements (McKenzie et al., 2001) generalized to a different domain. To this end, we used a visual identification task. At the yes–no stage, participants were presented with letter–number pairs and reported their confidence that a target letter and a target number appeared. The targets were conditionally independent. At the forced-choice stage, participants again reported confidence after being told (accurately) that the targets were mutually exclusive and exhaustive. Half of the participants were told to have their forced-choice confidence in the targets sum to 100%, and half were not. We expected that the most natural performance (and, hence, the best test of the models) would occur when participants were free to supply whatever confidence ratings they believed to be correct, regardless of whether they summed to 100%. Nevertheless, half of the participants were instructed to provide forced-choice confidence ratings that summed to 100% to see whether that would induce them to respond in more accordance with Equation 1, which requires that responses sum to 100%.

Method

Participants were 107 University of California, San Diego students who participated for course credit in undergraduate psychology courses. The experiment took place on a computer. During the yes–no phase, 64 letter–number pairs were presented sequentially. On each trial, one letter (*B, E,* or *R*) and one number (*5, 6,* or *8*) briefly appeared simultaneously side by side, separated by about 2 cm on the computer monitor. The target letter *B* occurred on half of the trials, and either *E* or *R* occurred on the other half. Similarly, the target number *6* occurred on half of the trials, and either *5* or *8* occurred on the other half. The left side of Table 1 shows the frequency of each letter–number pair comprising the 64 trials. After each trial, participants reported two numbers on a scale of 0–100: One was their confidence that *B* was present, and the other was their confidence that *6* was present. Zero corresponded to certainty that *B* (*6*) was absent, 50 corresponded to *B* (*6*) was equally likely present or absent, and 100 corresponded to certainty that *B* (*6*) was present. They were instructed to expect that *B* (*6*) would be present *X*% of the time when they reported *X*% confidence. They were discouraged from using 0, 50, and 100 except when they were truly certain or truly guessing. Half of the participants reported confidence in *B* first, and half reported confidence in *6* first. The letter appeared on the right and the number on the left on half of the trials, and the reverse was true for the other half. Letters and numbers were conditionally independent. For example, *B* was present on 50% of the trials regardless of whether *6* was present or absent.

Each trial began with a plus sign in the middle of the screen as a fixation point for 1,000 ms. The screen was then blank for 500 ms, followed by a letter–number pair. The pairs were present for about 30 ms on half of the trials and for about 50 ms on the other half. There were two speeds of presentation in order to get more variation in confidence reports. Immediately following the letter–number pair, a pound sign appeared in each of the two positions as a mask and was present for 1,000 ms. After reporting

[2] For the sake of completeness, it should be noted that in the algebraically equivalent Equation 5, c(*A*) and c(*B*) for a given pair define everything needed for computing predicted values of c(*A*,*B*) except for *s* (the posterior probability that the alternatives are mutually exclusive and exhaustive). However, unlike *r*, *s* is not constant across pairs. Instead, because it is a joint function of *r* and *q* (see Equation 4), *s* will differ for every pair (just as *q*, the prior probability, does). For that reason, *s* should not be thought of as a free parameter that can be estimated by adjusting its value to find the best fit. To produce an equation that can be fit to data, *s* in Equation 5 must be replaced by the right side of Equation 4 (this yields Equation 6).

Table 1
Presentations of Letter–Number Pairs at Each Stage in Experiments 1 and 2

Pair and speed	Yes–no	Forced choice	
		Experiments 1 and 2 (high probability)	Experiment 2 (low probability)
B/*5*			
Slow	4	4	4
Fast	4	4	4
B/*8*			
Slow	4	4	4
Fast	4	4	4
E/*6*			
Slow	4	4	4
Fast	4	4	4
R/*6*			
Slow	4	4	4
Fast	4	4	4
B/*6*			
Slow	8	0	2
Fast	8	0	2
E/*5*			
Slow	2	0	0
Fast	2	0	1
E/*8*			
Slow	2	0	1
Fast	2	0	0
R/*5*			
Slow	2	0	1
Fast	2	0	0
R/*8*			
Slow	2	0	0
Fast	2	0	1
Total	64	32	40

Note. For each letter–number pair with more than one presentation, the letter appeared on the left and the number on the right on 50% of the trials, and the reverse was true for the remaining 50%.

confidence that each of *B* and *6* was present, participants pressed the space bar to begin the next trial.

Before the 64 yes–no trials, there were 5 practice trials to familiarize participants with the task. They were instructed to seat themselves comfortably about 60 cm from the monitor and to maintain that distance during the entire task. The distance from the monitor, the spacing between the letter and number on the monitor, and the two speeds of presentation were chosen on the basis of pilot testing. The goal was to have a task in which participants were neither entirely certain of the target stimuli's presence nor purely guessing.

At the forced-choice stage, it was emphasized that, on every trial, either *B* or *6* would be present, but not both. Participants had to verbally acknowledge to the experimenter that they understood this information before beginning the forced-choice stage. Furthermore, half of the participants were instructed to have their confidence in *B* and *6* sum to 100, and half were not. Those who received such instructions had to verbally acknowledge that they understood this information as well before beginning. For this stage, only a subset of the pairs used in the yes–no stage was presented. Specifically, the pairs in which both *B* and *6* appeared and those in which neither *B* nor *6* appeared were eliminated, resulting in 32 trials (see the middle column of Table 1). Everything else was the same.

At the end of the experiment, participants were asked whether, when they were reporting confidence during the forced-choice stage, they had believed the instructions that either *B* or *6*, but not both, appeared on each trial. Because simply asking the question might influence whether they believed the information, it was emphasized that they were to report what they had believed when they were reporting their forced-choice confidence. They selected one statement from among three that best described their belief: One statement claimed full belief that either *B* or *6* appeared on every trial, one expressed some doubt, and one claimed not to believe the information.

Results

Of the 107 participants, 2 were eliminated because they reported 50% confidence on every trial, and 7 were eliminated for reasons given below in the *Individual-level results* section. As a manipulation check, we compared the uninstructed and instructed participants in terms of the extent to which their forced-choice confidence reports were additive (i.e., summed to 100). We calculated mean absolute deviation between 100 and summed confidence for each of the 32 forced-choice pairs for each participant. Greater means indicated greater nonadditivity (McKenzie, 1998; McKenzie et al., 2001). The uninstructed group's responses were less additive than those of the instructed group (Ms = 7.8 vs. 1.5) $t(96) = 4.61$, $p < .01$. Thus, the groups differed as expected.

Group-level results. We obtained the predictions of each model with a free parameter by fitting each model to the c(*A*,*B*) and c(*B*,*A*) values by adjusting the parameter (r or w) until the sum of squared deviations between the predicted and observed values was minimized. We used a quasi-Newton method minimization algorithm described by Fletcher (1972) to find the optimum value of each model's free parameter. Recall that the multiplicative and trust models are quantitatively virtually identical, but we present the variance accounted for in both cases just to reinforce that assertion.

For the uninstructed group, both the trust and multiplicative models accounted for 98.5% of the variance with best-fitting parameters of 0.85 and 0.67, respectively. That the trust model's r parameter was less than 1 indicates that, at the group level, uninstructed participants did not fully believe that the forced-choice alternatives were mutually exclusive and exhaustive. We expected r to fall between 0.5 and 1.0, but we did not constrain either model's parameter when fitting the models. The normative model (Equation 1) also performed quite well, accounting for 96.9% of the variance. Recall that the trust model with $r = 1$ is equivalent to the normative model (Equation 1). The trust model with $r = 0.85$ accounted for significantly more variance than the normative model, $F(1, 15) = 16.9$, $p < .05$.[3] That is, adding a free parameter accounted for a significantly greater percentage of the data variance than would be expected by chance alone. The same is true of the multiplicative model, of course (i.e., with $w = 0.67$,

[3] We report statistical tests on differences in variance accounted for between models at the group level only when one model is a more general form of the other. (The trust and multiplicative models are both general forms of Equation 1, the normative model.) We calculated F values by subtracting the general model's residual sum of squares (resulting from the least squares fit to the group data) from the normative model's residual sum of squares and dividing the difference by the general model's mean square error. The degrees of freedom are equal to the number of observations minus 1. These tests correct for differences in the number of free parameters between the models.

it also significantly outperformed Equation 1). Thus, these results replicate McKenzie et al.'s (2001) findings by showing that the normative model, relative to these other more general models, does not describe behavior as well.

A similar pattern of results occurred for the instructed group. The trust and multiplicative models accounted for 99.4% of the variance with respective best-fitting parameters of 0.84 and 0.64. The normative model accounted for 97.3% of the variance, significantly less than the other two models ($ps < .05$). Thus, the more general models outperformed the normative model even when confidence summed to (virtually) 100% for every forced-choice pair.

Recall that the normative model assumes that yes–no confidence in targets subsequently presented as forced-choice pairs is independent. In this experiment, because yes–no confidence in the target letter and in the target number was reported on the same trial, it is conceivable that reported confidence in one would influence reported confidence in the other. At the group level, however, we found only a very small positive correlation of .05. The mean individual-level correlation was .09 (the median was .07), which is small but significantly different from 0 ($p < .05$).

Individual-level results. We also fit the models at the individual level, resulting in a percentage of variance accounted for and (except for the normative model) a best-fitting parameter value for each model and participant. We eliminated 4 participants from the uninstructed group and 3 from the instructed group because their variance accounted for was negative for all models. (Negative variance accounted for occurs when the use of the best-fitting model results in greater error than the use of a constant—the mean—to predict confidence. Our models have no constants.)

For the uninstructed group, the trust and multiplicative models accounted for an average of 65.7% and 64.4% of the variance, respectively. Their respective average best-fitting parameters were 0.85 and 0.82. The normative model accounted for 57.7% of the variance on average. For the instructed group, variance accounted for was 66.1% and 65.9% for the trust and multiplicative models, and their respective parameters were 0.83 and 0.74. The normative model explained 58.3% of the variance. Thus, conclusions based on group-level analyses correspond to those based on individual-level analyses.

Believers versus skeptics. Most important, at the end of the experiment, participants were asked whether (a) they fully believed the information that the target items (*B* and *6*) during the forced-choice stage were mutually exclusive and exhaustive, (b) there was some doubt in their minds, or (c) they did not believe the information. For the uninstructed group, 30, 18, and 2 participants responded with (a), (b), and (c), respectively. For the instructed group, the corresponding numbers were 30, 15, and 3. Thus, about 60% of the participants reported fully believing the information that the targets were mutually exclusive and exhaustive during the forced-choice stage, and the additivity instructions had no effect, $\chi^2(2, N = 98) = 0.43$, $p = .81$. We categorized those participants who reported (a) as *believers* and those participants who reported either (b) or (c) as *skeptics.* The trust model implies that *r* should be higher for believers than it is for skeptics. The individual-level results are shown in Table 2. For the uninstructed group, *r*s were 0.94 and 0.80 for believers and skeptics, respectively, $t(48) = 3.11$, $p < .01$. (The *w* value for the multiplicative model was also lower for the skeptics, $p = .01$.) For the instructed group, *r*s were 0.92 and 0.78, respectively, $t(46) = 3.14$, $p < .01$; the *w* parameter was also significantly lower for skeptics ($p < .01$). Note that, as would be expected under the current account, the normative model did a pretty good job of explaining variance in believers' responses but did less well (relative to the other models) for skeptics' responses. Also of interest is that *r*, though high, was significantly less than 1 for believers in both the uninstructed, $t(29) = 2.36$, $p = .03$, and the instructed, $t(29) = 2.65$, $p = .01$, groups.

Discussion

The normative model (Equation 1) does not describe forced-choice behavior well, but the trust model, a new normative model that does not assume full belief that exactly one of the forced-choice items is true, does. Although both the trust and multiplicative models provided the best fit to the data, only the trust model appears conceptually accurate. The fact that almost 40% of the participants did not report full belief that the forced-choice items were mutually exclusive and exhaustive, despite efforts to get them to believe it (and that it was, in fact, true), is interesting and is itself evidence in favor of the trust model. Also important was that *r* was lower for those who reported disbelief that the forced-choice items were mutually exclusive and exhaustive, a straightforward prediction of the trust model. This was true for both the uninstructed and the instructed groups. The suboptimal impact of confidence in nonfocal alternatives during the forced-choice stage (relative to Equation 1), reported by McKenzie et al. (2001) and replicated

Table 2
Experiment 1: Individual-Level Results for Believers and Skeptics in Both the Uninstructed and Instructed Groups

	Uninstructed group				Instructed group			
	Believers		Skeptics		Believers		Skeptics	
Model	Parameter	VAF	Parameter	VAF	Parameter	VAF	Parameter	VAF
Trust	$r = 0.94$	65.4	$r = 0.80$	66.0	$r = 0.92$	73.6	$r = 0.78$	53.1
Multiplicative	$w = 0.96$	65.2	$w = 0.61$	63.1	$w = 0.89$	73.5	$w = 0.50$	52.7
Normative		62.1		51.0		69.0		40.7

Note. VAF = percentage of variance accounted for.

here, can be largely accounted for by the fact that some participants do not fully believe the experimenters' claim that the alternatives are mutually exclusive and exhaustive.

However, *r*, though close to 1, was significantly less than 1 for believers. Arguably, if skepticism that the forced-choice items are mutually exclusive and exhaustive is the reason for participants' failure to change confidence in accord with Equation 1, then those participants who did believe the stipulation should have behaved normatively—that is, *r* should have equaled 1 for them. If the trust model is correct, the most likely explanation for $r < 1$ for believers is that they (or at least a subset of them) did not fully believe the stipulation that the targets were mutually exclusive and exhaustive, although they reported otherwise. Alternatively, random error in participants' confidence reports might decrease *r*. McKenzie et al. (2001) reported simulation results showing that, even if the normative model (Equation 1) is correct (i.e., $w = 1$ in the multiplicative model), error in reported confidence will lower the multiplicative model's *w* to some extent.[4] A third possibility is that even participants who fully believe that the targets are mutually exclusive and exhaustive nonetheless underweight the strength of the nonfocal alternative.

Experiment 2

With respect to establishing the validity of the trust model, a shortcoming of Experiment 1 was that participants were sorted into believers and skeptics on the basis of postexperimental responses. Conceivably, participants who knew they were behaving appropriately (i.e., largely in accord with Equation 1) claimed to believe the experimenter, whereas those who knew they were not claimed to doubt the experimenter. If so, the participants' degree of belief in the experimenter's reliability did not determine their performance; instead, performance determined what they said about their degree of belief. One way to test the trust model further would be to use groups of participants expected to differ in terms of their degree of skepticism (e.g., those who have vs. those who have not been previously deceived in experiments). However, we thought it best to conduct an experiment in which participants were randomly assigned to conditions in order to avoid confounding variables. To do so, we used a surrogate of experimenter reliability. Specifically, some participants in Experiment 2 were told (accurately) that 100% of the forced-choice alternatives were mutually exclusive and exhaustive (as in Experiment 1), whereas others were told (accurately) that only 80% of the forced-choice alternatives were mutually exclusive and exhaustive. Thus, we assume that (truthfully) manipulating the base rate of mutually exclusive and exhaustive items in the forced-choice task is tantamount to manipulating experimenter reliability. We do not assume that participants fully believe either assertion by the experimenter, but because participants are randomly assigned, the degree of doubt, whatever it might be, should be the same for both conditions. Nevertheless, a participant in the 80% base rate condition (like a participant in Experiment 1 who does not trust the experimenter) should have greater doubt that the items are MEE than should a participant in the 100% base rate condition (like a participant in Experiment 1 who trusts the experimenter). We are, in essence, manipulating faith in the idea that the items are MEE rather than faith in the experimenter per se. Still, *r* should be sensitive to this manipulation.

Method

There were 103 participants from the same population used in Experiment 1. The procedure was identical to that used in Experiment 1 except that half of the participants (the low-probability group) were told (accurately) at the forced-choice stage that either *B* or *6*, but not both, would appear on 80% of the trials. On the remaining 20% of the trials, either both *B* and *6* would appear or neither *B* nor *6* would appear. The right side of Table 1 shows that 8 additional pairs were used in this condition during the forced-choice stage; 4 pairs included both *B* and *6*, and 4 pairs included neither. Thus, this group saw 40 pairs, and *B* and *6* were mutually exclusive and exhaustive in 32 of them (80%). The high-probability group was told (accurately), as in Experiment 1, that either *B* or *6*, but not both, would appear on all of the forced-choice trials. This group saw the 32 pairs listed in the middle column of Table 1. Participants in both groups had to verbally acknowledge to the experimenter that they understood the instructions before they could begin the forced-choice stage. Unlike in Experiment 1, no participants received additivity instructions.

Results

We eliminated 2 of the 103 participants (both in the high-probability group) because they were unable to see the letters or numbers, and 9 (4 high-probability and 5 low-probability) were eliminated because all three models resulted in negative variance accounted for. Data for the remaining 92 participants were analyzed.

We first note that, as in Experiment 1, there was only a small positive correlation (.07) between reported yes–no confidence in letter and number targets subsequently presented as forced-choice pairs at the group level. (Recall that the normative model assumes that c[*A*] and c[*B*] are independent.) The mean individual-level correlation was again small (0.08, equal to the median) but significantly different from 0 ($p < .05$).

Group-level results. The left side of Table 3 shows the results for the high-probability group (which is analogous to the unstructured group in Experiment 1). The trust and multiplicative

[4] We conducted simulations to examine whether error in confidence reports alone can account for $r < 1$. We conducted them using c(*A*) and c(*B*) values that were varied factorially between 0.1 and 0.9 in steps of 0.1. For each of the resulting 81 c(*A*) and c(*B*) pairs, we calculated c(*A*,*B*) using the trust model with $r = 1$. However, each c(*A*), c(*B*), and c(*A*,*B*) value was disturbed by random error. In particular, rather than using the "true" value, we drew a value from a beta distribution with the same mean value. Enough error was introduced that fitting the trust model to the data for each of 20 simulated participants accounted for about 65% of the variance on average, which was the mean variance accounted for in Experiment 1 at the individual level. The mean *r* value across simulated participants was generally less than 1.0 (15 simulated participants had values less than 1.0, and 5 had values greater than 1.0), but not by much. The estimated values of *r* ranged from 0.94 to 1.03, with a mean of 0.98. Thus, these simulations do not suggest that the somewhat lower values of *r* that we obtained in Experiment 1 for self-reported believers were the result of error alone. Nevertheless, it is conceivable that larger deviations from 1.0 could result from error alone under other conditions, for example, modeling error with something other than the beta distribution or using different distributions of c(*A*) and c(*B*).

models again performed well, accounting for 97.8% of the variance. These two models also performed well in the low-probability condition, as shown on the right side of Table 3. (We fitted the models using the same 32 trials for both probability groups.) The trust and multiplicative models accounted for more variance than did the normative model (Equation 1) in both conditions (ps $<$ .05)—that is, for both models, adding a free parameter accounted for more variance than would be expected on the basis of chance alone. Note that, as one would expect, the normative model performed especially poorly in the low-probability condition (presumably due, in large part, to the fact that it is not the normative model when less than 100% of the forced-choice items are mutually exclusive and exhaustive). It is important to note that the trust model's r parameter was much higher for the high-probability group than it was for the low-probability group. As can be seen, the multiplicative model's w parameter was higher as well.

Individual-level results. Table 4 shows the mean results after fitting the models to the data for each participant.[5] All of the trends from the group-level analyses are seen here as well. Both the multiplicative and trust models outperformed the normative model, but they were similar to each other in terms of goodness of fit. Also, the normative model provided an especially poor fit in the low-probability condition. Most important is that the trust model's r parameter was higher for the high-probability group than it was for the low-probability group, $t(90) = 3.32$, $p = .001$. This was also true for the multiplicative model's w parameter ($p = .01$).

Discussion

Manipulating the experimenter's (accurate) assertion regarding the objective probability that the forced-choice targets were mutually exclusive and exhaustive affected the trust model's r parameter as expected: Lowering the probability from 100% to 80% (i.e., lowering participants' belief that the items are mutually exclusive and exhaustive) lowered the parameter value. This finding reinforces our claim that the trust model not only provides a good fit to the data but is conceptually accurate as well.

General Discussion

Deviations from the purported normative model prescribing change in confidence between yes–no and forced-choice tasks appear to be explained best by the fact that participants do not fully believe a key task parameter, namely, that the forced-choice items are mutually exclusive and exhaustive. The good fit provided by the trust model (Equation 6), a new normative model that does not assume full belief in the task parameter, indicates that participants are largely responding optimally, given their degree of belief that the forced-choice items are mutually exclusive and exhaustive. The trust model's only free parameter, r, which corresponds to the perceived reliability of the experimenter when he or she is asserting that the forced-choice items are mutually exclusive and exhaustive, was lower in Experiment 1 for participants who expressed some skepticism, rather than none, regarding the assertion. It is not obvious how changes in the multiplicative model's parameter could be explained. This suggests that the trust model is not only quantitatively accurate but is conceptually accurate as well. (The fact that 40% of participants in Experiment 1 expressed at least some disbelief in the task parameter is interesting and provides additional evidence for the trust model.) Experiment 2, in which we used random assignment to conditions that differed objectively in terms of the probability that the forced-choice options were mutually exclusive and exhaustive, showed that the trust model continued to provide a good fit, and r again differed between the groups as expected.

A limitation of our experiments is that we did not experimentally manipulate the reliability of the experimenter (which is what r theoretically captures). In Experiment 1, we divided participants into believers and nonbelievers after the fact. The value of r differed for these two groups in the expected way, and our model would have been seriously challenged had that result not occurred, but the lack of random assignment left open the possibility that some other difference between the two groups was responsible for the effect on r. In Experiment 2, we did use random assignment, but we manipulated a surrogate of experimenter reliability. Again, the value of r changed in the predicted direction, and had that not happened, the trust model would not be viable. Although the results of Experiments 1 and 2 are sufficient to drive home our main point (namely, that it is a mistake to declare the behavior of participants to be nonnormative without taking into account their degree of belief in task parameters), an attempt to manipulate the characteristics of the experimenter that influence perceived reliability would be an important further test of our model.

Table 3
Experiment 2: Mean Results for the Group-Level Analyses

	High-probability group		Low-probability group	
Model	Parameter	VAF	Parameter	VAF
Trust	$r = 0.76$	97.8	$r = 0.61$	98.1
Multiplicative	$w = 0.49$	97.8	$w = 0.19$	98.0
Normative		92.7		81.0

Note. VAF = percentage of variance accounted for.

Table 4
Experiment 2: Mean Results for the Individual-Level Analyses

	High-probability group		Low-probability group	
Model	Parameter	VAF	Parameter	VAF
Trust	$r = 0.81$	68.0	$r = 0.68$	66.9
Multiplicative	$w = 0.60$	67.6	$w = 0.39$	65.6
Normative		52.4		41.5

Note. VAF = percentage of variance accounted for.

[5] One high-probability participant's multiplicative w value was 25.1, and this extreme outlier was eliminated from the analysis. In addition, one other high-probability participant could not be fit by the multiplicative and normative models because of division by zero, which occurs for these models when c(A) and c(B) are both 0 or both 1.

The trust model is, of course, specific to tasks in which confidence is expressed in alternatives whose degree of (in)dependence has changed. However, deriving and testing the trust model is only an example of a general approach that we believe has considerably larger implications. The idea that participants may not fully believe key task parameters is not one that is generally taken into consideration in experiments designed to assess whether participants behave in a normative manner. Often, researchers arrive at the conclusion that participants do not behave normatively, just as McKenzie et al. (2001) did. Not only does our proposed approach make salient various assumptions experimenters might otherwise take for granted, but it also highlights the fact that there are often multiple normatively defensible responses to a given situation (e.g., Birnbaum, 1983; Einhorn & Hogarth, 1981; Gigerenzer, 1991; Hilton, 1995; McKenzie, in press-a, in press-b; McKenzie & Mikkelsen, in press; Oaksford & Chater, 1994; Schwarz, 1996; Sher & McKenzie, 2003).

To illustrate how the approach could be applied to a different area of research, consider again the base-rate studies discussed earlier. Recall that Gigerenzer et al. (1988) found that increasing the believability of the random sampling procedure led to responses that were much more normative. The following is a new normative model of the Bayesian odds that a hypothesis (H) is true given data (D), but that also takes into account the judge's degree of belief (r) that the data are randomly sampled from a population with a certain base rate, $p(\mathrm{H})$:

$$\frac{p(\mathrm{H}|\mathrm{D})}{p(\sim\mathrm{H}|\mathrm{D})} = \frac{p(\mathrm{D}|\mathrm{H})\{r[p(\mathrm{H})] + (1 - r)(0.5)\}}{p(\mathrm{D}|\sim\mathrm{H})\{r[1 - p(\mathrm{H})] + (1 - r)(0.5)\}},$$

where $0 \leq r \leq 1$. When $r = 1$, the equation reduces to the odds form of Bayes's theorem, the traditional normative response. When $r = 0$, uniform base rates are assumed (which is tantamount to ignoring the presented base rates). This model is normative under the assumption that the judge does not fully believe that the data are randomly sampled and shows that underweighting base rates can be rational if $r < 1$. Such a model has the potential for answering questions regarding whether participants are making errors in Bayesian tasks or are responding reasonably, given their (understandable) skepticism about the random sampling stipulation.

It is not hard to find examples of claims about normative errors relying on participants' full belief in key task parameters. Another example that, like the one above, regards verbal assertions about sampling comes from Hamill, Wilson, and Nisbett (1980). These authors showed participants a staged videotaped interview with a prison guard who was either humane or inhumane. Some participants were told that the guard they saw was atypical, some were told that the guard was typical, and some were told nothing about typicality. They then answered questions about the characteristics of prison guards in general. The main finding was that the assertions about typicality had no effect. For example, if participants had seen the humane guard, they reported that prison guards were generally fair regardless of whether they were told that the guard was typical or atypical. The researchers interpreted the lack of effect as showing participants' insensitivity to sample bias. The interpretation, however, relied entirely on participants believing the experimenters' claim about typicality. It is possible that participants simply did not believe the claim (which, of course, had to be untrue for at least some of the participants).

We mention just one more example. Ross, Lepper, and Hubbard (1975) presented participants with pairs of suicide notes, one of which was said to be authentic and one of which was said to be inauthentic. Participants were to judge which note was authentic for a series of such pairs and, after each judgment, received feedback. All participants received predetermined (i.e., false) feedback, indicating either that most of their judgments were correct or that most were incorrect. During the initial debriefing, it was explained to participants that they had been randomly assigned to either a positive- or a negative-feedback condition, and that hence, the feedback was independent of their performance. They were then asked to provide judgments of how well they would do on an additional series of trials. The main finding was that participants given positive feedback predicted that they would do better than did those given negative feedback. The authors argued that this "belief perseverance" was normatively unjustified because the initial basis for their perception of their ability had been "completely discredited" (p. 880). However, in this conclusion, the authors assume that participants fully believed the experimenter when told that their feedback was predetermined. Note the participants' quandary after being told they had been deceived ("Were they lying to me then, or are they lying to me now?"), and any skepticism about what the experimenter said about the false feedback leads to results that the authors consider irrational. One might argue that participants should fully believe what they are told during debriefing, and that therefore there should be no effect of the initial feedback. Whatever the merits of this argument, we note one more aspect of this experiment: The initial debriefing was also part of the experiment, and participants were deceived then, too. It was only during the final debriefing that participants were told the true purpose of the experiment. Maybe the only irrational thing to do in any experiment is to fully believe anything the experimenter tells you.

In short, although we think it is a good idea to make important task parameters believable, it is probably not reasonable to assume that participants fully believe the parameters even under these circumstances. In Experiment 1, we emphasized (accurately) to participants that the forced-choice alternatives were mutually exclusive and exhaustive, and they had to verbally acknowledge to the experimenter that they understood this before performing the task. Despite these efforts, almost 40% of the participants later reported that they doubted that the forced-choice items were mutually exclusive and exhaustive. We believe it is best to accept participant skepticism as an important—and tractable—variable in laboratory experiments, especially those that compare behavior to a normative standard. The success of the trust model in the present context shows that it is both desirable and feasible to develop normative models in which it is not assumed that participants believe key assumptions that are often taken for granted by experimenters.

References

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47,* 1597–1611.

Asch, S. E. (1955). Opinions and social pressure. *Scientific American, 193,* 31–35.

Bender, R. H. (1998). Judgment and response processes across two knowledge domains. *Organizational Behavior and Human Decision Processes, 75,* 222–257.

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology, 96,* 85–94.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology, 32,* 53–88.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Hillsdale, NJ: Erlbaum.

Ferrel, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Decision Processes, 26,* 32–53.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90,* 239–260.

Fletcher, R. (1972, June). *FORTRAN subroutines for minimization by quasi-Newton methods.* (Report No. R7125). Harwell, England: United Kingdom Atomic Energy Authority.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." *European Review of Social Psychology, 2,* 83–115.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 513–525.

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 21–31.

Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: Generalizing from atypical cases. *Journal of Personality and Social Psychology, 39,* 578–589.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24,* 383–451.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118,* 248–271.

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

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology: I* (pp. 103–189). New York: Wiley.

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology, 26,* 209–239.

McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 771–792.

McKenzie, C. R. M. (1999). (Non)complementary updating of belief in two hypotheses. *Memory & Cognition, 27,* 152–165.

McKenzie, C. R. M. (in press-a). Framing effects in inference tasks—and why they're normatively defensible. *Memory & Cognition.*

McKenzie, C. R. M. (in press-b). Judgment and decision making. In K. Lamberts & R. L. Goldstone (Eds.), *Handbook of cognition.* London: Sage.

McKenzie, C. R. M., & Mikkelsen, L. A. (in press). A Bayesian view of covariation assessment. *Cognitive Psychology.*

McKenzie, C. R. M., Wixted, J. T., Noelle, D. C., & Gyurjyan, G. (2001). Relation between confidence in yes/no and forced-choice tasks. *Journal of Experimental Psychology: General, 130,* 140–155.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371–378.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631.

Ortmann, A., & Hertwig, R. (1997). Is deception acceptable? *American Psychologist, 52,* 746–747.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32,* 880–892.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation.* Mahwah, NJ: Erlbaum.

Sher, S., & McKenzie, C. R. M. (2003). *Information leakage from logically equivalent frames.* Manuscript submitted for publication.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 681–690.

(*Appendixes follow*)

Appendix A

Derivation of the Trust Model

In this appendix, we derive the trust model for combining confidence judgments when there is some intermediate amount of trust in the claim that the two alternatives at hand are mutually exclusive and exhaustive. The trust model is first presented in Equation 5. Here we will show that this model is normative given a quantified amount of uncertainty concerning the claim that the two alternatives are mutually exclusive and exhaustive.

Let A be the event that the first of a given pair of possibilities is true, and let B be the independent event that the second is true. Let X be the event that A and B are mutually exclusive and exhaustive—that exactly one of the two possibilities is true. Formally, we can write X as $[(A \wedge \bar{B}) \vee (\bar{A} \wedge B)]$, where the $\wedge$ symbol is logical "and", the $\vee$ symbol is logical "or", and the covering horizontal bar indicates logical negation. Let T be the event that a teacher (e.g., the experimenter) indicates that A and B are mutually exclusive and exhaustive—that X is true. Note that, due to lack of trust in the teacher's statement, T may be true while X is false, or X may just so happen to be true of a given pair of options, even if the teacher did not specifically dictate exclusiveness and exhaustiveness. Let $P(\cdot)$ be an appropriate probability mass function over these events. Given this kind of uncertainty, the normative expression of $c(A,B)$ is $P(A|T)$. This quantity may be computed as follows:

$$P(A|T) = P(A \wedge T)/P(T) = \frac{P(A \wedge T \wedge X) + P(A \wedge T \wedge \bar{X})}{P(T)}$$

$$= \frac{P(A|T \wedge X)P(X|T)P(T) + P(A|T \wedge \bar{X})P(\bar{X}|T)P(T)}{P(T)}$$

$$= P(A|T \wedge X)P(X|T) + P(A|T \wedge \bar{X})P(\bar{X}|T).$$

At this point, we must make the assumption that A and T are independent when conditioned on X. In other words, we must assume that the teacher's instructions provide no new information about the truth of A once we already know whether X (the mutual exclusivity and exhaustiveness dictate) is actually true. If we already know that exactly one of A and B is true, the teacher's proclamation of this fact changes nothing. This assumption reduces the above expression to:

$$P(A|T) = P(A|X)P(X|T) + P(A|\bar{X})P(\bar{X}|T)$$

$$= P(A|X)P(X|T) + P(A|\bar{X})[1 - P(X|T)].$$

This equation contains a term, $P(X|T)$, that expresses the trust assigned to the teacher's proclamation. This term is the degree of belief in the mutual exclusivity and exhaustiveness dictate, given that the teacher provided it. Note that this term is equivalent to the parameter s introduced in Equation 4. Substituting in s and manipulating this expression results in:

$$P(A|T) = sP(A|X) + (1 - s)P(A|\bar{X})$$

$$= \frac{sP(A|X)P(\bar{X}) + (1 - s)P(A|\bar{X})P(\bar{X})}{P(\bar{X})}$$

$$= \frac{sP(A|X)P(\bar{X}) + (1 - s)P(A|\bar{X})P(\bar{X})}{P(\bar{X})}$$

$$+ \frac{(1 - s)P(A|X)P(X) - (1 - s)P(A|X)P(X)}{P(\bar{X})}$$

$$= \frac{\begin{array}{l}P(A|X)[sP(\bar{X}) - (1 - s)P(X)] \\ + (1 - s)[P(A|X)P(X) + P(A|\bar{X})P(\bar{X})]\end{array}}{P(\bar{X})}$$

$$= \frac{P(A|X)[s - sP(X) - P(X) + sP(X)] + (1 - s)P(A)}{P(\bar{X})}$$

$$= P(A|X)\left[\frac{s - P(X)}{1 - P(X)}\right] + P(A)\left[\frac{1 - s}{1 - P(X)}\right]$$

$$= P(A|X)\left[\frac{s - P(X)}{1 - P(X)}\right] + P(A)\left\{\frac{1 - P(X) - [s - P(X)]}{1 - P(X)}\right\}$$

$$= P(A|X)\left[\frac{s - P(X)}{1 - P(X)}\right] + P(A)\left[1 - \frac{s - P(X)}{1 - P(X)}\right].$$

Noting that $P(X)$ is the same as the parameter q, introduced in Equation 3, and defining the weighting parameter w to be equal to $(s - q)/(1 - q)$ one may write this expression in terms of w. Finally, by recognizing that $P(A|X)$ is the normative probability of A when mutual exclusivity and exhaustiveness is known with certainty—a probability that is calculated by the normative model shown in Equation 1—one may rewrite this expression as:

$$P(A|T) = wP(A|X) + (1 - w)P(A)$$

$$= w\frac{P(A)[1 - P(B)]}{P(A)[1 - P(B)] + P(B)[1 - P(A)]} + (1 - w)P(A).$$

This expression for $P(A|T)$ is entirely in terms of the priors on A and B, along with the weighting parameter, w. It is equivalent to Equation 5, demonstrating that this expression provides the Bayes optimal way to combine estimates of the truth of A and B given a mutual exclusivity and exhaustiveness proclamation that is trusted to some limited degree, quantified by the parameter w.

Appendix B

Reparameterizing the Trust Model

Although Equation 5 provides a normative method for updating beliefs in the face of information concerning the mutual exclusivity and exhaustiveness of two options, it has one major drawback: Its weighting parameter, *w*, is sensitive to the prior beliefs in *A* and *B*. We would prefer a formulation of this model that includes a single free parameter that may be seen as independent of the priors on *A* and *B*, capturing only information about the general reliability of the teacher (e.g., the experimenter). Fortunately, such a reparameterization of the trust model exists. It is derived below, using the notation from Appendix A.

It is assumed that the judge retains information about the reliability of the teacher in terms of two probabilities: a false-positive rate, $P(T|\bar{X})$, and a false-negative rate, $P(\bar{T}|X) = 1 - P(T|X)$. These two probabilities are combined into a single trust parameter, called *r:*

$$r = \frac{P(T|X)}{P(T|X) + P(T|\bar{X})}.$$

In this appendix, we show how the trust model can be expressed in terms of *r*.

Initially identifying a few simple relationships between the probabilities of interest facilitates the reparameterization process. First, note that the contributions to *r* of the false-positive rate and the false-negative rate are symmetric:

$$\frac{P(T|\bar{X})}{P(T|X) + P(T|\bar{X})} = \frac{P(T|X) + P(T|\bar{X}) - P(T|X)}{P(T|X) + P(T|\bar{X})}$$

$$= 1 - \frac{P(T|X)}{P(T|X) + P(T|\bar{X})} = 1 - r. \quad \text{(B1)}$$

Next, note that the denominator of the reparameterized trust model, shown in Equation 6, has the following simplified form:

$$\frac{P(T)}{P(T|X) + P(T|\bar{X})} = \frac{P(T|X)P(X) + P(T|\bar{X})P(\bar{X})}{P(T|X) + P(T|\bar{X})}$$

$$= rP(X) + (1 - r)P(\bar{X})$$

$$= r\{P(A)[1 - P(B)] + P(B)[1 - P(A)]\}$$

$$+ (1 - r)\{P(A)P(B) + [1 - P(A)][1 - P(B)]\}. \quad \text{(B2)}$$

One further relationship is worth noting before proceeding with the derivation of the reparameterized trust model:

$$\frac{P(T) - P(T|X)}{1 - P(X)} = \frac{P(T|X)P(X) + P(T|\bar{X})P(\bar{X}) - P(T|X)}{1 - P(X)}$$

$$= \frac{P(T|\bar{X})[1 - P(X)] - P(T|X)[1 - P(X)]}{1 - P(X)}$$

$$= P(T|\bar{X}) - P(T|X). \quad \text{(B3)}$$

With these three identities in hand, we can begin to algebraically manipulate the original trust model (Equation 5):

$$P(A|T) = w\frac{P(A)[1 - P(B)]}{P(A)[1 - P(B)] + P(B)[1 - P(A)]} + (1 - w)P(A)$$

$$= \left(\frac{s - q}{1 - q}\right)\left\{\frac{P(A)[1 - P(B)]}{P(X)}\right\} + \left(1 - \frac{s - q}{1 - q}\right)P(A)$$

$$= P(A)\left(\frac{s - q}{1 - q}\right)\left[\frac{1 - P(B)}{P(X)}\right] + P(A)\left(\frac{1 - s}{1 - q}\right)\left(\frac{P(X)}{P(X)}\right)$$

$$= \frac{P(A)}{P(X)(1 - q)}\left\{\left[\frac{P(T|X)P(X)}{P(T)} - P(X)\right][1 - P(B)]\right.$$

$$\left. + P(X)\left[1 - \frac{P(T|X)P(X)}{P(T)}\right]\right\}$$

$$= \frac{P(A)}{1 - q}\left\{\left[\frac{P(T|X)}{P(T)} - 1\right][1 - P(B)] + \left[1 - \frac{P(T|X)q}{P(T)}\right]\right\}$$

$$= \frac{P(A)}{1 - q}\left\{\frac{P(T|X)}{P(T)}[1 - P(B) - q] + P(B)\right\}$$

$$= \frac{P(A)}{1 - q}\left\{\frac{P(T|X)}{P(T)}(1 - q) + P(B)\left[1 - \frac{P(T|X)}{P(T)}\right]\right\}$$

$$= \frac{P(A)}{P(T)(1 - q)}\{P(T|X)(1 - q) + P(B)[P(T) - P(T|X)]\}$$

$$= \frac{P(A)}{P(T)}\left[P(T|X) + P(B)\frac{P(T) - P(T|X)}{1 - P(X)}\right].$$

We may now substitute in Equation B3, resulting in

$$P(A|T) = \frac{P(A)}{P(T)}\{P(T|X) + P(B)[P(T|\bar{X}) - P(T|X)]\}$$

$$= \frac{P(A)}{P(T)}\{P(T|X)[1 - P(B)] + P(T|\bar{X})P(B)\}$$

$$= \frac{P(A)}{P(T)/[P(T|X) + P(T|\bar{X})]}\left\{\frac{P(T|X)}{P(T|X) + P(T|\bar{X})}[1 - P(B)]\right.$$

$$\left. + \frac{P(T|\bar{X})}{P(T|X) + P(T|\bar{X})}P(B)\right\}.$$

Substituting in the definition of *r* and the results of Equations B1 and B2 gives us

$$P(A|T) = \frac{P(A)\{r[1 - P(B)] + (1 - r)P(B)\}}{r\{P(A)[1 - P(B)] + P(B)[1 - P(A)]\} + (1 - r)\{P(A)P(B) + [1 - P(A)][1 - P(B)]\}}$$

$$= \frac{rP(A)[1 - P(B)] + (1 - r)P(A)P(B)}{r\{P(A)[1 - P(B)] + P(B)[1 - P(A)]\} + (1 - r)\{P(A)P(B) + [1 - P(A)][1 - P(B)]\}}.$$

This is the reparameterized trust model as it appears in Equation 6. This equation is formally equivalent to the original trust model, though this version is parameterized in terms of *r*, which does not vary with the priors on *A* and *B*, given our assumptions.

Received May 19, 2003
Revision received February 3, 2004
Accepted February 10, 2004 ■