Journal of Experimental Psychology: General
2001, Vol. 130, No. 1, 140–155


0096-3445/01/$5.00 DOI: 10.1037//0096-3445.130.1.140

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

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

David C. Noelle
Carnegie Mellon University

Gohar Gyurjyan
University of California, San Diego

Yes–no and forced-choice tasks are common in psychology, but the empirical relation between reported confidence in the 2 tasks has been unclear. The authors examined this relation with 2 experiments. The general experimental method had participants first report confidence in the truth of each of many general knowledge statements (a yes–no task) then report confidence in them again when the statements were put into pairs where it was known that one statement was true and one was false (a forced-choice task). At issue was how confidence in the statements changed between the yes–no task and the forced-choice task. Two models, including the normative one, were ruled out as descriptive models. A linear model and a multiplicative model remain viable contenders.

Imagine taking a true–false test consisting of 50 statements, but in addition to simply reporting whether each statement is true or false, you also report your subjective probability, or confidence, that each is true. Assume that your confidence in the truth of two particular statements, A and B, is 80% and 40%, respectively. Next, imagine a subsequent test in which the same 50 statements are arranged into 25 pairs where one statement in each pair is true and one is false. Your task now, in addition to selecting the statement you think is the true one, is to report confidence in each statement, with the knowledge that exactly one statement in each pair is true. A and B form one such pair. How confident would you now be that A is true? That B is true? How confident should you be?

These two tests capture the essence of a yes–no and a forced-choice task, respectively, both of which are used extensively in psychology. For example, in a typical experiment on recognition memory, participants study a list of items (such as words) and then take a recognition test in which they attempt to discriminate targets (words that appeared on the list) from lures (words that did not). The recognition test usually involves either a yes–no or a forced-choice format. In the former, one item is presented at a time, and participants must decide whether it is a target or a lure (see, e.g., Glanzer & Adams, 1990; Roediger & McDermott, 1995). In the latter, items are presented in pairs consisting of one target and one lure, and the task is to choose the target (see, e.g., Glanzer & Bowles, 1976; Kim & Glanzer, 1993). In both tasks, participants are often asked to supply a confidence rating for each recognition decision (Glanzer, Adams, Iverson, & Kim, 1993; Ratcliff, McKoon, & Tindall, 1994; Roediger & McDermott, 1995; Stretch & Wixted, 1998). Similar examples are easy to find in other areas. In studies of perception, yes–no tasks sometimes involve reporting whether a stimulus is present, and forced-choice tasks involve reporting which one of two (or more) stimuli is present (see, e.g., Creelman & Macmillan, 1979; Swets, Markowitz, & Franzen, 1969). In categorization experiments, participants might be asked whether an object belongs to a particular category or asked which of two categories it belongs to (for examples of the former, see Hampton, 1979, 1998; McCloskey & Glucksberg, 1978; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; for examples of the latter, see Gluck & Bower, 1988; Goldstone, 1996). In the area of judgment and decision making, participants sometimes report confidence in the truth of individual statements (e.g., "absinthe is a liqueur"; Fischhoff, Slovic, & Lichtenstein, 1977; Wallsten & González-Vallejo, 1994), and sometimes there are multiple alternatives ("absinthe is (a) a liqueur or (b) a precious stone"; Fischhoff et al., 1977; Koriat, Lichtenstein, & Fischhoff, 1980). Both tasks are not only easy to find examples of but also are virtually always used in tandem. If one of the tasks is being used to study a particular area or phenomenon, the other is also likely being used.

Analogues to the two tasks can be found outside the laboratory as well. In our opening example, the yes–no task is essentially a true–false test, and the forced-choice task is essentially a multiple-choice test. As another example, imagine a physician assessing the likelihood that a patient has each of two illnesses, A and B, that occur in the population independently. That is, knowing that a patient has A says nothing about whether the patient has B. Under these conditions, the physician might assess the probability of A and B to be 80% and 40%, respectively. Additional information then becomes known, leading the physician to believe that the

Craig R. M. McKenzie, John T. Wixted, and Gohar Gyurjyan, Department of Psychology, University of California, San Diego; David C. Noelle, Center for the Neural Basis of Cognition, Carnegie Mellon University.

This research was supported by National Science Foundation Grant SBR-9515030.

We thank Tom Wallsten and Mike Ziolkowski for their many constructive criticisms.

Correspondence concerning this article should be addressed to Craig R. M. McKenzie, Department of Psychology, University of California, San Diego, La Jolla, California 92093-0109. Electronic mail may be sent to cmckenzie@ucsd.edu.

patient has either A or B, but not both. The task has changed from yes–no (for each of A and B) to forced choice (A vs. B). How confident would (and should) the physician now be that the patient has illness A?

To clarify how we conceive of the relation between yes–no and forced-choice tasks, as well as illustrate the normative response, consider the medical diagnosis scenario in more detail. A and B are assumed to be statistically independent, $p(A) = 0.8$, and $p(B) = 0.4$. These probabilities (and their complements) are represented by the marginal probabilities in Figure 1. If there were 100 patients, 80 would be expected to have A, and 40 would be expected to have B. Furthermore, each patient would belong to one of four categories: Those who have (a) both A and B, (b) A but not B, (c) B but not A, or (d) neither A nor B. The independence assumption allows for easily calculating the expected number of patients in each category. One need only multiply the probabilities of the respective events. For Category a, $p(A) = 0.8$, $p(B) = 0.4$, and their product is 0.32. Thus, 32 of the 100 patients would be expected to belong to the A&B category (Category a), and the corresponding (joint) probability is shown in the upper left cell in Figure 1. Analogous calculations show that the expected number of patients in the other three categories is 48, 8, and 12, respectively.

Now assume that the patient is known to belong to either the A&~B category or the ~A&B category. That is, the patient has either A or B, but not both. Belonging to the A&B category or to the ~A&~B category is no longer a possibility. How confident should the physician be that the patient has illness A (i.e., belongs to the A&~B category)? There are 56 patients in the A&~B and ~A&B categories combined, and 48 of them belong to A&~B. Hence, the probability is 48/56, or about 0.86. Similarly, the probability that the patient belongs to the ~A&B category is 8/56, or about 0.14. In our running example, then, confidence in A should increase from 80% to 86%, and confidence in B should decrease from 40% to 14%. The normative model is derived in the Appendix.

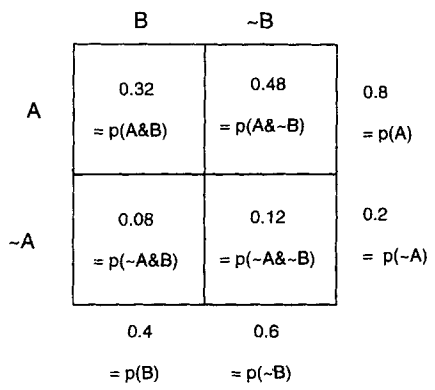


*Figure 1.* The medical diagnosis scenario (as described in the text) in 2 × 2 form. The marginal probabilities correspond to the probability of illness A and illness B (and their complements) in a yes–no task. Patients can belong to one of four categories: They can have (a) both A and B (upper left cell), (b) A but not B (upper right), (c) B but not A (lower left), or (d) neither A nor B (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, a patient can belong to only the A&~B category or the ~A&B category: The patient has either A or B, but not both. The other two categories—A&B and ~A&~B—are no longer possible. 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 $0.08/(0.08 + 0.48) = 14\%$.

Thus, we view a yes–no task involving A and B to be one in which all four above categories, or outcomes, are possible, at least theoretically. For the sake of simplicity, we assume that A and B are statistically independent, but our approach is not limited by this assumption. We believe that the independence assumption holds for many yes–no tasks, but there are undoubtedly cases where knowing that A is true would influence, to some degree, belief in B. Accordingly, in the Appendix, we present the normative model without the independence assumption as well.

Furthermore, we view the prototypical forced-choice task involving A and B to be one in which participants understand that the A&B and ~A&~B categories have been eliminated, leaving only A&~B and ~A&B. Again, this seems to capture many forced-choice tasks, but there are situations where A and B are not mutually exclusive and exhaustive. Our discussion assumes, though, that A and B are independent in the yes–no task and are mutually exclusive and exhaustive in the forced-choice task. More formally, we assume that $p(A|B) = p(A)$ and $p(B|A) = p(B)$ in the former task and that $p(A|B) = p(B|A) = 0$ in the latter.

To recap, despite widespread use of yes–no and forced-choice tasks, surprisingly little is known about how confidence between them is related. Furthermore, intuition—ours at least—does not provide a clear guide. In our running example, we do not think it is obvious, despite the above normative analysis, that confidence in A will increase in the forced-choice task relative to the yes–no task. Indeed, we present and test three plausible psychological models, one of which predicts an increase in confidence in A under such circumstances, and two of which predict a decrease.

Much previous research has examined how confidence changes on the basis of new information (see, e.g., Anderson, 1981; Hogarth & Einhorn, 1992). However, in such belief-updating tasks, the new information consists of new evidence, and the competing hypotheses do not change. In contrast, we are interested in situations where the new information is not new evidence in the usual sense but information that two hypotheses formerly considered independent are mutually exclusive and exhaustive. Furthermore, the fact that the hypotheses change to mutually exclusive and exhaustive also changes the set of competing hypotheses. When A and B are independent, A's competitor is not-A, and B's competitor is not-B. When A and B are mutually exclusive and exhaustive, they compete directly with each other.

Research more closely linked to our topic was conducted by Van Wallendael and Hastie (1990; Robinson & Hastie, 1985; Van Wallendael, 1989). Suspects in a murder mystery were used as mutually exclusive and exhaustive hypotheses, and various clues were presented. After each clue, participants updated their confidence that each suspect was guilty. The authors occasionally added or dropped a suspect and examined how confidence changed. Thus, the new information occasionally consisted of a change in the hypothesis set, much like our current concern. However, the hypotheses were always mutually exclusive, which resulted in a fundamentally different problem than the one we examine. Furthermore, the authors' primary dependent measure was the sum of confidence in the competing hypotheses, whereas we investigate formal models for adjusting confidence.

The article is organized as follows. In the first part, we present the normative model and three descriptive models of how confi-

dence might change between yes–no and forced-choice tasks. In the second part, we report two experiments indicating that two plausible models—including the normative model—can be ruled out. Subsequently, we report results from computer simulations that examine the effect of error in confidence judgments on the models' performance. We then discuss the relationship between the normative model and the signal-detection interpretation of confidence. In the final section, we summarize our contributions and suggest future research.

## Three Models for Changing Confidence Between Yes–No and Forced-Choice Tasks

Consider two statements, or hypotheses, A and B, that are independent. For example, one might be asked for confidence in the truth of each of the following statements:

A. The U.S. population is greater than 250 million.

B. Plato was born before Socrates.

Under such conditions, knowing that, for example, A is true would have no effect on confidence that B is true, which is what we mean when we say that A and B are independent. Let $c(A)$ and $c(B)$ represent confidence in A and B, respectively, when A and B are independent, and let $0 < c(A), c(B) < 1$. We use the term confidence to denote how often the judge expects to be correct in the long run. In our experiments, participants were instructed to expect to be correct $x\%$ of the time when reporting $x\%$ confidence, allowing us to make normative claims about how confidence should change. Assume that $c(A) = 0.8$ and $c(B) = 0.4$ and that it becomes known that A and B are mutually exclusive and exhaustive: One of A and B is true, and one is false. How confident would one now be that A is true? That B is true? We denote confidence in A under these conditions by $c(A,B)$, indicating confidence in A, given that B is the sole alternative. Similarly, $c(B,A)$ corresponds to confidence in B, with A as its sole alternative. Following are three descriptive models of $c(A,B)$ as a function of $c(A)$ and $c(B)$.

### *Multiplicative Model*

The normative model, demonstrated above, and derived in the Appendix, is the following:

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

To calculate $c(B,A)$, simply switch $c(A)$ and $c(B)$ in Equation 1. An important property of this model, relative to the others we present below, is that $c(A,B)$ is always more extreme than $c(A)$ when $c(A)$ and $c(B)$ are on different sides of 0.5. As we showed earlier, for example, if $c(A)$ and $c(B)$ are respectively 0.8 and 0.4, $c(A,B) = 0.86$ and $c(B,A) = 0.14$. If $c(A)$ is greater than 0.5, then $c(A,B)$ ranges from $c(A)$ to 1 as $c(B)$ ranges from 0.5 to 0. If $c(A)$ is less than 0.5, then $c(A,B)$ ranges from $c(A)$ to 0 as $c(B)$ ranges from 0.5 to 1.

The multiplicative model we tested is a more general model with one free parameter:

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

The parameter, $w$, is associated with each term involving $c(B)$ and determines the extent to which $c(B)$ affects $c(A,B)$ (holding $c[A]$ and $c[B]$ constant). We expect $w$ to vary between 0 and 1. When $w = 1$, Equations 1 and 2 are identical. 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 then equals $c(A)$. 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 (see, e.g., Evans, 1989; Fischhoff & Beyth-Marom, 1983; Klayman & Ha, 1987; McKenzie, 1994, 1998).

Because A and B in the forced-choice task are mutually exclusive and exhaustive, $c(A,B)$ and $c(B,A)$ should (normatively) sum to 1 (i.e., be additive). However, if $w < 1$, then confidence is not additive (except in the special case where confidence in the yes–no task, $c[A]$ and $c[B]$, happens to sum to 1). Additivity generally decreases as $w$ moves from 1. If $w = 0$ in our running example, $c(A,B) + c(B,A) = 0.8 + 0.4 = 1.2$, indicating superadditivity. More generally, if $c(A) + c(B)$ exceeds 1 and $w < 1$, confidence in the forced-choice task is superadditive. Similarly, if $c(A) + c(B)$ is less than 1 and $w < 1$, then confidence in the forced-choice task is subadditive (i.e., sums to less than 1). The relationship between underweighting the alternative and nonadditivity has been studied empirically (McKenzie, 1998, 1999).

We often refer to $c(A,B)$ rather than to both $c(A,B)$ and $c(B,A)$ for simplicity's sake, even though participants reported both values. Because of this, we usually refer to B as the alternative. However, it should be kept in mind that A is the alternative when reporting $c(B,A)$.

Note that the normative model (Equation 1) and the multiplicative model (Equation 2) can be written in a more convenient way if the probabilities are expressed as odds. The relationship between odds (O) and a corresponding probability ($p$) is $O = p/(1 - p)$. Thus, $O_A = c(A)/(1 - c[A])$ and $O_B = c(B)/(1 - c[B])$. Substituting these values into Equation 1 and rearranging terms yields

$$O_{A,B} = O_A/O_B, \quad (3)$$

which, in the more general form analogous to Equation 2, is

$$O_{A,B} = O_A/(O_B)^w. \quad (4)$$

### *Linear Model*

Another way that confidence could change is that the algebraic difference between confidence in A and B in the yes–no task might remain the same in the forced-choice task but sum to 1 in the latter task. The following equation captures such a process:

$$c(A,B) = 0.5(c[A]) + 0.5(1 - c[B]). \quad (5)$$

In words, $c(A,B)$ is the mean of $c(A)$ and the complement of $c(B)$. If $c(A)$ were 0.8 and $c(B)$ were 0.4, $c(A,B)$ would equal to 0.7, and $c(B,A)$ would equal 0.3. The algebraic difference between $c(A)$ and $c(B)$ and between $c(A,B)$ and $c(B,A)$ is 0.4, but the latter values sum to 1. Part of the motivation behind testing such a model is that linear models often predict behavior well, even in tasks where nonlinear or multiplicative judgments might be expected (see, e.g., Dawes, 1979; Hoffman, 1960). Brehmer (1980) suggested that people's intuitive strategies might often be linear because linear models are simple and perform well under a variety of circumstances.

Note that the linear model led to a decrease in $c$(A,B) relative to $c$(A) in our running example, a result opposite to that of the normative and multiplicative models. For the linear model, whenever $c$(A) and $c$(B) sum to greater than 1, the resulting forced-choice confidence values decrease. The model essentially takes half the difference between 1 and the sum of $c$(A) and $c$(B) and subtracts that amount from $c$(A) and $c$(B) to arrive at $c$(A,B) and $c$(B,A). When $c$(A) and $c$(B) sum to less than 1, forced-choice confidence increases in an analogous manner. Both $c$(A,B) and $c$(B,A) either decrease or increase together relative to their yes–no counterparts, and by the same amount. This behavior is different from that of the normative and multiplicative models, where $c$(A,B) might increase relative to $c$(A) whereas $c$(B,A) decreases relative to $c$(B).

The linear model we tested is a generalization of Equation 5 with one free parameter. It allows the coefficients to differ from 0.5 but constrains them to sum to 1:

$$c(\mathrm{A,B}) = (1 - w/2)(c[\mathrm{A}]) + (w/2)(1 - c[\mathrm{B}]), \quad (6)$$

and we expect $w$ to vary between 0 and 1. When $w = 1$, note that Equations 5 and 6 are equivalent. The smaller $w$ is, the smaller the impact that $c$(B) has on $c$(A,B), holding $c$(A) and $c$(B) constant. If $w = 0$, then $c$(B) has no impact, and $c$(A,B) = $c$(A). As with the multiplicative model, one can interpret $w$ in terms of the extent to which the strength of the alternative is taken into account. In addition, when $w < 1$, nonadditivity generally occurs: If $c$(A) + $c$(B) is less than 1, $c$(A,B) + $c$(B,A) is less than 1, and if $c$(A) + $c$(B) is greater than 1, $c$(A,B) + $c$(B,A) is greater than 1. Despite the general nonadditivity when $w < 1$, the difference between $c$(A,B) and $c$(B,A) always equals the difference between $c$(A) and $c$(B).

### *Ratio Model*

The simple form of our third model holds constant the ratio, rather than the difference, between $c$(A) and $c$(B) and between $c$(A,B) and $c$(B,A), while forcing the latter values to sum to 1:

$$c(\mathrm{A,B}) = c(\mathrm{A})/[c(\mathrm{A}) + c(\mathrm{B})]. \quad (7)$$

If $c$(A) = 0.8 and $c$(B) = 0.4, then $c$(A,B) = 0.667, and $c$(B,A) = 0.333. The 2-to-1 ratio of confidence in the yes–no task is maintained in the forced-choice task, but confidence in the latter task sums to 1. Equation 7 normalizes confidence in A and B. The psychological appeal of this model is simply that the relative confidence in A and B is held constant. If one were twice as confident in A as in B in the yes–no task, then, according to this model, one would be twice as confident in A as in B in the forced-choice task. Similar models have been proposed by Tversky and Koehler (1994; see also Koehler, 1996) in the context of strength of evidence judgments and by Luce (1959) in the context of choice frequencies.

Note that, like the linear model but unlike the multiplicative model, $c$(A,B) decreased relative to $c$(A) in our running example. The ratio model adjusts both $c$(A,B) and $c$(B,A) in the same direction, just like the linear model. Relative to $c$(A) and $c$(B), $c$(A,B) and $c$(B,A) both increase (when $c$[A] + $c$[B] < 1) or both decrease (when $c$[A] + $c$[B] > 1). Unlike the linear model, however, $c$(A,B) and $c$(B,A) do not change by the same amount. Instead, change is proportional to confidence in the yes–no task. In the above example, $c$(A,B) decreased 0.133 relative to $c$(A), whereas $c$(B,A) decreased 0.067 relative to $c$(B). The first change is twice the second because $c$(A) is twice $c$(B).

The ratio model we tested is a more general form of Equation 7:

$$c(\mathrm{A,B}) = c(\mathrm{A})/[c(\mathrm{A}) + c(\mathrm{B})]^w. \quad (8)$$

We expect $w$ to vary between 0 and 1 and again interpret it as the extent to which the strength of the alternative, $c$(B), is taken into account or affects $c$(A,B). When $w = 1$, Equations 7 and 8 are identical. At the other extreme, where $w = 0$, $c$(B) has no impact on $c$(A,B), which would equal $c$(A). The smaller $w$ is, the smaller the impact that $c$(B) has on $c$(A,B), holding $c$(A) and $c$(B) constant. Furthermore, to the extent that $w$ differs from 1, additivity generally decreases (in the same manner as the linear model), but Equation 8 nonetheless holds constant the ratio between $c$(A) and $c$(B) when producing $c$(A,B) and $c$(B,A).

To summarize, we have presented a normative model (Equation 1) and three plausible descriptive models (Equations 2, 6, and 8) for adjusting confidence in two hypotheses when they change from independent to mutually exclusive and exhaustive. The descriptive models are presented in Table 1 for reference. Each descriptive model has one free parameter, facilitating comparisons of performance. Furthermore, the free parameter in each model has a straightforward psychological interpretation.

## Overview of Experiments

It is easiest to provide an overview of our two experiments by discussing the second one first. In the first part of Experiment 2, participants reported confidence on a scale of 0 to 100 in the truth of 60 general knowledge statements. Participants were told that half the statements were true and half were false. In the second part of Experiment 2, the same 60 statements were presented in 30 pairs, where it was known that one statement in each pair was true and one was false. For each pair of statements, A and B, in the second part, reported confidence in the first part corresponded to $c$(A) and $c$(B), and reported confidence in the second part corresponded to $c$(A,B) and $c$(B,A). At issue was how confidence in each statement changed between the first part (a yes–no task) and the second part (a forced-choice task).

Unlike Experiment 2, Experiment 1 controlled participants' confidence in the first part (i.e., $c$[A] and $c$[B]). This allowed for a wide range of $c$(A) and $c$(B) values that facilitated model testing and eliminated the large between-participants variability in confidence that is typical for general knowledge statements. Rather than reporting $c$(A) and $c$(B), participants were given predetermined

Table 1
*Three Descriptive Models*

| Model | Equation |
|---|---|
| Multiplicative | $c(\mathrm{A,B}) = c(\mathrm{A})[1 - c(\mathrm{B})]^w / [c(\mathrm{A})(1 - c[\mathrm{B}])^w + c(\mathrm{B})^w(1 - c[\mathrm{A}])]$ |
| Linear | $c(\mathrm{A,B}) = (1 - w/2)(c[\mathrm{A}]) + w/2(1 - c[\mathrm{B}])$ |
| Ratio | $c(\mathrm{A,B}) = c(\mathrm{A})/[c(\mathrm{A}) + c(\mathrm{B})]^w$ |

*Note.* $c$(A) and $c$(B) represent confidence at the yes–no stage (in A and B, respectively). $c$(A,B) represents confidence in A when A and B are forced-choice alternatives. $w$ is a free parameter and determines the extent to which $c$(B) affects $c$(A,B).

values and told to imagine that these corresponded to their confidence in the truth of various individual (unseen) general knowledge statements. These confidence values were arranged in pairs such that one corresponding unseen statement was true and one was false. For example, one pair was 90/50, indicating (hypothetically) 90% confidence that the first unseen statement was true when presented individually and 50% confidence in the second one. Participants then updated the two values given that it was now known that one of the unseen statements was true and one was false.

In both experiments, half the participants were told that $c$(A,B) and $c$(B,A) must sum to 100% for each pair, and half were given no such instructions. We included this manipulation because recent studies have revealed that subjective probabilities can be nonadditive, even for two mutually exclusive and exhaustive hypotheses (McKenzie, 1998, 1999). As our discussion of the models indicates, underweighting the strength of the nonfocal alternative leads to nonadditive judgments. On the other hand, people do often report additive probabilities, and in some cases (e.g., in decision analysis or in the laboratory), confidence judgments in mutually exclusive and exhaustive hypotheses are forced to sum to 100%. Thus, there is reason to believe that the instructions will affect additivity, and testing the models under only the additive or the nonadditive condition would seem to us incomplete. It is possible that the best model for describing behavior depends on whether or not confidence is additive.

## Experiment 1

### *Method*

Participants were 78 students at the University of California at San Diego who received partial credit for introductory psychology courses. They were given a two-page booklet, the first page of which provided instructions. Participants were to imagine that they had previously been presented with 50 individual statements, such as "The population of the U.S. is greater than 200 million" and "Socrates was born before Sophocles," and that they had reported how confident they were in the truth of each. They were to think of their reported confidence in terms of long-run frequencies. For example, they were to expect 90% of the statements they had reported 90% confidence in to be true. They were then to imagine that the 50 individual statements had been arranged in 25 pairs. In each pair, one statement was true, and one was false. It was pointed out that the two statements above could be such a pair because one was true and one was false.

The second page presented 25 pairs of values, labeled A and B. Each value was to be considered their confidence in the truth of an individual statement. Participants did not see any actual statements, just their confidence. They were to report new confidence (between 1 and 99) in each statement, A and B, given that one was true and one was false. As with confidence in the original statements, they were to expect $x$% of the statements in which they reported $x$% confidence to be true.

Five possible values of confidence were used for each statement in each pair: 10, 30, 50, 70, and 90. Every combination of these values was used, resulting in 25 pairs. Half the participants were presented with the pairs in a predetermined random order, and half were presented with the reverse order. Furthermore, half the participants were told that their confidence in A and B for each pair should sum to 100, whereas half were not.

### *Results*

One outlier was eliminated from the group instructed to have $c$(A,B) and $c$(B,A) sum to 100, and one was eliminated from the group given no such instructions, leaving 39 and 37 participants, respectively. Criteria for exclusion are given below, where we discuss individual-level analyses.

*Group-level analyses.* We first checked the distribution of the $c$(A,B) and $c$(B,A) responses. Not surprisingly, responses tended to fall on salient values on the 99-point scale: 67.0% and 74.9% of the responses were either 1, 10, 20, 30, . . . , 90, or 99 for the uninstructed and instructed groups, respectively (see also, e.g., Fischhoff et al., 1977). We also checked for a difference in additivity between the two groups. As is usually done, we calculated the sum of $c$(A,B) and $c$(B,A) for each pair. The mean sum was 101.8 for the uninstructed group and 100.1 for the instructed group, $t(74) = 1.61$, $p > .10$, implying no difference in additivity. Furthermore, neither value differs significantly from 100 (both $ps > .10$). These results are typical, and the usual conclusion is that confidence is additive when there are two mutually exclusive and exhaustive hypotheses (Robinson & Hastie, 1985; Teigen, 1983; Tversky & Koehler, 1994; Van Wallendael, 1989; Van Wallendael & Hastie, 1990; Wallsten, Budescu, & Zwick, 1993). However, McKenzie (1998) argued that mean summed confidence is at least sometimes inappropriate for measuring additivity. This might appear paradoxical, but consider the following extreme possibility: One could be 0% confident in each alternative composing one pair and 100% confident in each alternative in another pair. Mean summed confidence across the two pairs is 100%, but it would be misleading to regard these confidence reports as additive. (Mean summed confidence does, however, reveal whether there is an overall bias toward superadditivity or subadditivity.) A more appropriate measure under the current circumstances is to calculate, for each pair, the absolute deviation between 100 and the sum of confidence in the two alternatives. The greater the mean absolute deviation across the 25 pairs, the greater the nonadditivity. This measure revealed a clear difference between the groups, with the uninstructed group exhibiting a larger mean absolute deviation than the instructed group: $M$s = 15.9 versus 0.2, $t(74) = 9.6$, $p < .001$. Thus, the groups differed in a meaningful way for testing our models.

The left side of Figure 2 shows the results for each of the three models for the uninstructed group. The top left graph shows the relation between mean reported confidence (divided by 100) and the predictions of the multiplicative model. The predictions were obtained through fitting the multiplicative model to the $c$(A,B) and $c$(B,A) values by adjusting $w$ until the sum of squared deviations between the predicted and observed values was minimized. A quasi-Newton method minimization algorithm described by Fletcher (1972) was used to find the optimum value of $w$. There are 50 data points in the panel, corresponding to the 50 mean $c$(A,B) and $c$(B,A) values (25 pairs) reported by participants, plotted against the model's predictions. The identity line indicates where all the data points would lie if the model predicted participants' confidence perfectly. The panel also indicates the best-fitting $w$ parameter, 0.54, for the least squares fit of the multiplicative model. That $w$ is less than 1 indicates that $c$(A) had a larger impact than $c$(B) when reporting $c$(A,B)—and that $c$(B) had a larger impact than $c$(A) when reporting $c$(B,A)—relative to the normative model. We expected $w$ to vary between 0 and 1, but we did not constrain the parameter to this range when fitting the models. Listed below $w$ is the percentage of variance accounted for by the multiplicative model. The value of 95.7 indicates that the multiplicative model fit the data quite well.

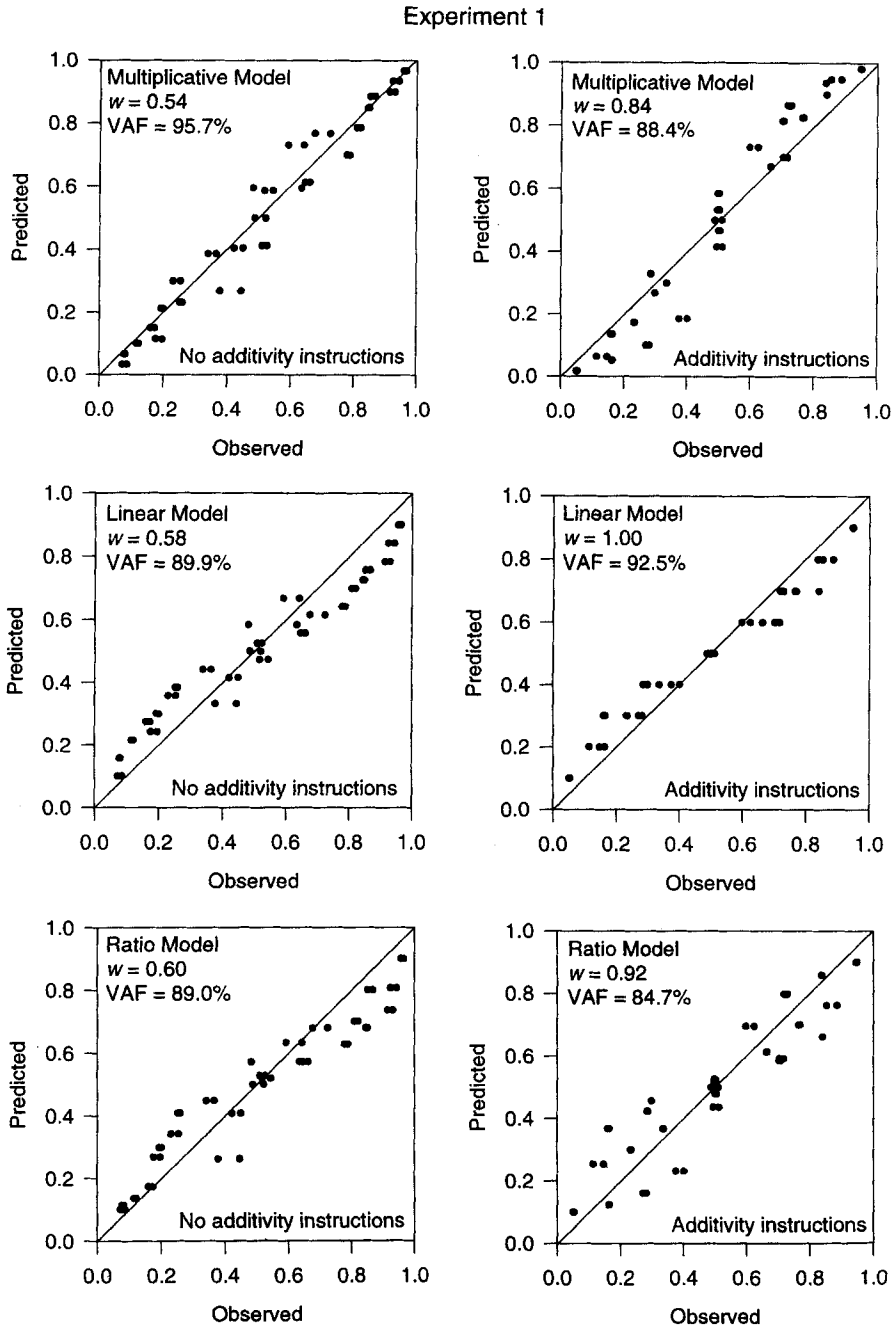


*Figure 2.* Experiment 1: Each of the three models' performance for the group without additivity instructions (left side) and for the group with additivity instructions (right side). Plotted is participants' mean confidence against each model's predictions. Predictions were obtained through fitting each model to the reported $c$(A,B) and $c$(B,A) values by adjusting $w$ until the sum of squared deviations between the predicted and observed values was minimized. The resulting $w$ and variance accounted for (VAF) are shown in each panel. $w$ = the degree to which forced-choice confidence was affected by confidence in the nonfocal alternative.

Recall that the multiplicative model with $w = 1$ is equivalent to the normative model. We found that the multiplicative model (with $w = 0.54$) accounted for significantly more variance than the normative model: $F(1, 49) = 133.4$, $p < .05$.

The middle left panel in Figure 2 shows the linear model's results for the uninstructed group. According to this model, confidence in the alternative was underweighted (because $w < 1$). Relative to the multiplicative model, variance accounted for was somewhat worse.

Note also that, unlike the multiplicative model, the linear model tended to overpredict low confidence values and underpredict high ones. That is, judgments tended to be more extreme than the linear

model predicted. This systematic bias might lead one to think that the model is fundamentally wrong, but, as we discuss later in the computer simulation section, bias is misleading if one makes the reasonable assumption that there is random error in participants' reported confidence. Error makes the linear and ratio models biased even if they are correct. In contrast, the simulations show that, for the models and conditions investigated here, variance accounted for is useful for determining which model is correct even in the presence of nonnormal random error. Accordingly, though we sometimes mention bias when reporting results, we emphasize variance accounted for when determining which models can and cannot be rejected.

The ratio model, shown in the bottom left panel, accounted for slightly less variance than the linear model, and considerably less than the multiplicative model. The $w$ of 0.60 indicates that the alternative was underweighted. Note that the bias is in the same direction as that of the linear model: Confidence was more extreme than predicted by the model.

The three panels on the right side of Figure 2 show the results for the group given the additivity instructions. As can be seen, the linear model accounted for the most variance, although judgments were again more extreme than predicted by the model. Note also that $w = 1$, indicating that the instructions led participants to weight $c$(A) and $c$(B) equally, according to the linear model. Surprisingly, this means that the best fitting model for this group is an equal-weighted linear model (Equation 5). The other two models still resulted in $w$s considerably less than 1, though. Furthermore, the multiplicative model's bias was in the direction opposite to the other two: The model tended to underpredict low confidence and overpredict high confidence. The multiplicative model again accounted for significantly more variance than the normative model ($p < .05$).

Comparing results between the groups, it can be seen that each mode's $w$ moved closer to 1 with additivity instructions, indicating that $c$(B) had a larger impact on $c$(A,B), and that $c$(A) had a larger impact on $c$(B,A). The additivity instructions led the linear model to account for more variance and led the multiplicative and ratio models to account for less variance.

*Individual-level analyses.* We also fit the models at the individual level, resulting in a percentage of variance accounted for and a $w$ for each participant. Participants were eliminated if variance accounted for was negative for all three models. (Negative variance accounted for occurs when the best fitting model results in greater error than simply using a constant—the mean—to predict confidence. Our models have no constants.) As mentioned, this occurred for 2 (out of 78) participants, one in each group. These occurrences were rare and extreme, and we have few qualms about eliminating them for present purposes.

Table 2 shows the mean $w$ and variance accounted for across the individual-level analyses for each model and group. Consider first the uninstructed group, shown on the left side of the table. The multiplicative model accounted for significantly more variance than both the linear and ratio models (both $t$s $> 4.3$, $p$s $< .001$), which did not differ from each other ($p = .29$). Also of interest is that, out of the 111 $w$ values for this group (37 participants $\times$ 3 models), only 8 were outside the expected 0–1 range, with 5 less than 0 and 3 greater than 1. (One can interpret $w > 1$ as overweighting the nonfocal alternative.) Aside from one participant who had a relatively large negative $w$ for all three models, the largest departure from the 0–1 interval was only 0.06.

Table 2
*Experiment 1: Mean Results for the Individual-Level Analyses*

| | Uninstructed group | | Instructed group | |
|---|---|---|---|---|
| Model | $w$ | VAF | $w$ | VAF |
| Multiplicative | 0.56 | 80.9 | 0.83 | 72.5 |
| Linear | 0.58 | 75.0 | 1.00 | 80.7 |
| Ratio | 0.59 | 74.4 | 0.91 | 73.1 |

*Note.* VAF corresponds to percentage of variance accounted for; $w$ = the degree to which forced-choice confidence was affected by confidence in the nonfocal alternative.

The right side of Table 2 shows the results for the instructed group. The linear model accounted for more variance than both the ratio and multiplicative models ($p$s $< .015$), which did not differ from each other ($t < 1$). In addition, of the 117 $w$ values in this group (39 participants $\times$ 3 models), only 6 fell outside the expected 0–1 range. All 6 values were greater than 1, but the largest was only 1.08. Also of interest is that the $w$ value for the linear model was 1.00 for 33 of the 39 participants in this group.

Relative to the group-level analyses (Figure 2), these individual-level analyses reveal less variance accounted for, but rank order in terms of performance largely remained the same. Note also that each mean $w$ in Table 2 is virtually identical for each model and group to its counterpart in the group-level analyses.

Table 3 shows for each group the percentage of participants whom each model fit best as measured by variance accounted for. We included these analyses because, though the models might differ by only a small amount in terms of mean variance accounted for, one model could nonetheless provide a better fit for every participant. The first column of numbers reveals that, for the uninstructed group, the multiplicative model accounted for the most variance for 81.1% of the participants. For the instructed group, Table 3 indicates that the linear model was the winner. These results again largely reflect the group-level analyses.

*Qualitative analyses.* There are four data points that provide critical tests between the multiplicative model and the other two. For the 70/10 and 10/70 pairs presented to participants, the multiplicative model predicts that new confidence corresponding to 10 will decrease, whereas the linear and ratio models predict an increase. For the 90/30 and 30/90 pairs, the multiplicative model predicts that new confidence corresponding to 90 will increase, whereas the linear and ratio models predicts a decrease. (These critical tests arise when $c$[A] and $c$[B] are on different sides of 50 and sum to less than or greater than 100, as in our 80/40 running example.) At the group level, the multiplicative prediction was correct all four times for the uninstructed group, and the linear/ratio prediction was correct all four times for the instructed group. Individual-level results revealed the same pattern: Uninstructed participants reported changes in confidence on the critical items in a manner consistent with the multiplicative model 2.7 times out of 4, on average, whereas the instructed participants did so only 1.1 times, $t(74) = 5.7$, $p < .001$. These qualitative findings are consistent with the model-fitting results in that the multiplicative model described the uninstructed group best and the linear model described the instructed group best.

Table 3
*Experiment 1: Percentage of Participants in Each Group for Whom Each Model Accounted for the Most Variance*

| Model | Uninstructed group | Instructed group |
|---|---|---|
| Multiplicative | 81.1 | 35.9 |
| Linear | 16.2 | 64.1 |
| Ratio | 2.7 | 0.0 |

### *Discussion*

The results of Experiment 1 allow us to effectively rule out two otherwise plausible models. First, the normative model, with no free parameter, accounted for significantly less variance than the multiplicative model, where the best fitting free parameter was less than 1. (The models are equivalent when $w = 1$ in the multiplicative model.) In particular, confidence in the nonfocal alternative had suboptimal impact in the forced-choice task. Second, the ratio model was consistently weak, accounting for relatively low variance across groups. Virtually no participants were fit best by this model.

The multiplicative model accounted for the most variance for the uninstructed group, but the linear model accounted for the most variance for the instructed group. These results were consistent across group- and individual-level analyses. They were also consistent with analyses based on items where the two models made qualitatively different predictions. The qualitative findings imply that the linear model is not doing well merely because it is robust (i.e., able to account for high variance even if the model is wrong). Instructions appear to have influenced the process used to update confidence. Also of interest is that, for the instructed group, the best fitting $w$ parameter was 1 for the linear model, indicating that an equal-weighted linear model performed best in that group (Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975).

The additivity instructions affected not only additivity and which model fit best but the $w$ parameter as well. Each model's parameter moved closer to 1 with the instructions, as expected. Furthermore, the actual range of the $w$ parameter in the individual-level analyses was largely within the expected range for each model and group. These findings lend credence to the validity of the parameter in each model.

Our data indicate that the multiplicative and linear models cannot be ruled out, which is not to say that either model is correct. Other models with the same number of free parameters may fit the data as well, perhaps better. Distinguishing between descriptively accurate models is nontrivial because even models with the same number of free parameters can differ in their inherent flexibility (Myung & Pitt, 1997). Thus, conservatively, the contribution of Experiment 1 is to help reduce the field of viable candidates. Because both the multiplicative and linear models are conceptually simple and descriptively accurate—much more accurate than the normative model, for example—they seem worthy of further consideration.

Arguably, however, it is premature to completely reject the alternative models because Experiment 1 was largely hypothetical for participants. They did not report confidence in the yes–no task, only the forced-choice task. Indeed, participants did not see any hypotheses or statements, only numbers. It could be that the relation between confidence in yes–no and forced-choice tasks differs when the judgments are inspired by true, rather than hypothetical, uncertainty. In Experiment 2, participants reported confidence in the truth of actual general knowledge statements presented in both yes–no and forced-choice format.

## Experiment 2

### *Method*

Participants were 100 students from the same population as in Experiment 1. The experiment took place on computer. In the first part, participants reported confidence in the truth of 60 randomly ordered general knowledge statements that covered five different area: the arts, history, science, sports, and geography. There were approximately equal numbers of statements from each category. A range of difficulty was sought, but the statements were not representatively sampled (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994). Participants were told that half the statements were true and half were false. An example of a true literature statement is "Jonathan Swift wrote *Gulliver's Travels*," and an example of a false history statement is "Thomas Jefferson was the second U.S. president." As in Experiment 1, reported confidence was defined in terms of long-run accuracy. Participants were to expect $x$% of the statements in which they reported $x$% confidence to be true.

In the second part, the 60 statements were arranged in 30 pairs. For each pair, one statement was true and one was false, and participants were told this. This information was also present on the screen the entire second part. Participants were presented with pairs of statements labeled A and B, and they reported confidence in the two statements contiguously. All participants saw the same 30 pairs, though in different random orders. Statements within a pair were from different content areas, with only a few exceptions. Participants were not shown their previously reported confidence in the individual statements. As in Experiment 1, half the participants were instructed to have their confidence for each pair of statements sum to 100, and half were not.

### *Results*

On the basis of the same criteria used in Experiment 1, we eliminated one participant in the uninstructed group and 2 in the instructed group, leaving 49 and 48 participants, respectively. In addition, participants were allowed to respond with values between 0 and 100 in this experiment, but 0 and 100 were recoded to 1 and 99 because some models we tested cannot accommodate certainty.

*Group-level analyses.* Even more so than in Experiment 1, responses tended to fall on the salient scale values: 87.8% and 91.4% of the uninstructed and instructed groups' $c$(A,B) and $c$(B,A) responses were either 1, 10, 20, . . . , 90, or 99. We again checked to see if instructions affected additivity. As before, mean summed confidence did not differ between the uninstructed and instructed groups, $M$s = 102.7 and 100.4, respectively, $t(95) = 1.70$, $p = .09$; but mean absolute deviation between 100 and summed confidence was different, $M$s = 6.9 vs. 2.1, respectively, $t(95) = 3.18$, $p = .002$. The effect of instructions was weaker than in Experiment 1, where the uninstructed group was less additive than its current counterpart and the instructed group was more additive.

The three panels on the left side of Figure 3 show the group-level results (predicted vs. observed values) for the uninstructed group. The multiplicative model accounted for the most variance (91.1%). That $w$ was less than 1 for this model indicates that $c$(B) had less impact on $c$(A,B)—and that $c$(A) had less impact on $c$(B,A)—than prescribed by the normative model. The ratio model

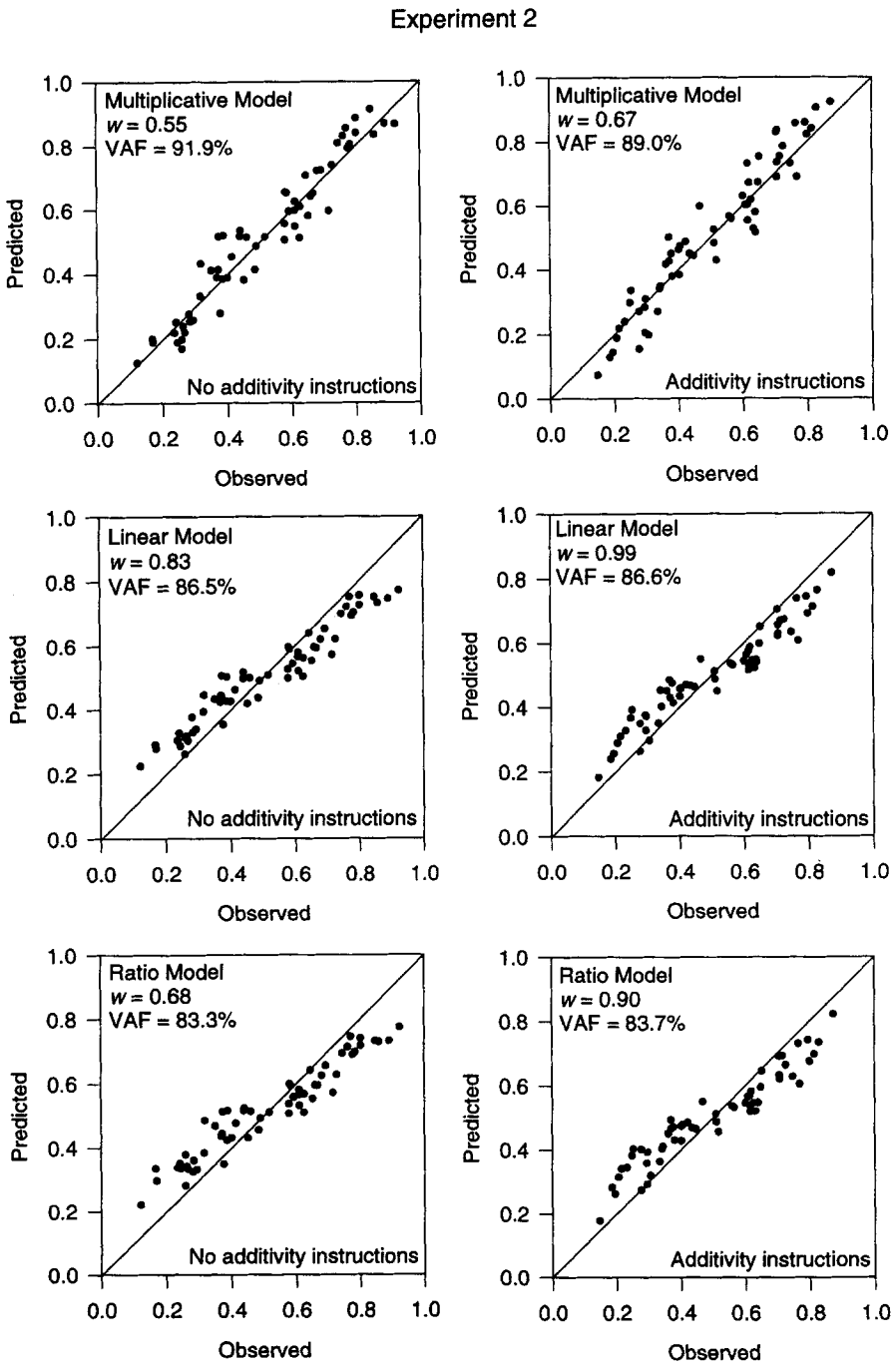


*Figure 3.* Experiment 2: The models' performance for the group without additivity instructions (left side) and for the group with additivity instructions (right side). Plotted is participants' mean confidence against each model's predictions. VAF = variance accounted for; $w$ = the degree to which forced-choice confidence was affected by confidence in the nonfocal alternative.

accounted for the least variance. Though not shown in Figure 3, the normative model accounted for significantly less variance than the multiplicative model ($p < .05$).

The right side of Figure 3 tells a similar story for the instructed group. The multiplicative model accounted for the most variance and the ratio model the least. As in Experiment 1, the additivity instructions increased $w$ in the three models. The linear model's $w$ was again very close to 1 for the instructed group, indicating that, according to this model, $c$(A) and $c$(B) played roughly equal roles in determining $c$(A,B) and $c$(B,A). The other two models indicate otherwise, however. The normative model again accounted for significantly less variance than the multiplicative model ($p < .05$).

Consistent with the relatively small effect of instructions on additivity, differences in the models' performance between the

groups were small, though the directions of the effects were largely consistent with those of Experiment 1. The multiplicative model accounted for less variance with the additivity instructions, whereas the linear and ratio models accounted for slightly more variance.

*Individual-level analyses.* The mean results of the individual-level analyses are shown in Table 4. For the uninstructed group, the ratio model accounted for less variance than both the linear and multiplicative models ($ts > 3, ps < .01$), which did not differ from each other ($t < 1$). Of the 147 $w$ values for this group (49 participants × 3 models), 13 fell outside the 0–1 range (11 were greater than 1), but only 5 values were more than 0.05 outside the interval.

For the instructed group, the linear model accounted for more variance than the other two models ($ps < .05$), which did not differ from each other ($t < 1$). Of the 144 $w$ values, only 7 fell outside the 0–1 interval. All 7 were greater than 1, but the largest was 1.04.

Relative to the group level, the individual-level analyses revealed lower variance accounted for. In terms of the models' performance relative to each other, the only systematic difference across level of analysis was that the linear model performed better at the individual level for both groups. Whereas, at the group level, the multiplicative model accounted for the most variance for both groups, it tied the linear model at the individual level for the uninstructed group and was outperformed by the linear model for the instructed group.

Table 5 shows the percentage of participants whom each model fit best according to variance accounted for. For the uninstructed group, the linear and multiplicative models each fit about half the participants best, and for the instructed group, the linear model did somewhat better in terms of variance accounted for. The ratio model's performance was again weak.

*Qualitative analyses.* Qualitative tests were more complicated in this experiment because participants were not supplied with $c$(A) and $c$(B) as in Experiment 1. At the group level, there were eight critical items in each group (coincidentally). That is, there were eight items where confidence in the yes–no stage led the multiplicative model to make a qualitatively different prediction from the linear and ratio models. The required conditions were that $c$(A) and $c$(B) be on different sides of 0.5 and sum to less than 0.9 or more than 1.1. (Confidence had to sum to considerably less than 1 or more than 1 because the linear and ratio models predict no change in confidence to the extent that the sum of $c$[A] and $c$[B] is 1.) Of the eight critical items in each of the uninstructed and the instructed groups at the group level, four were consonant with the multiplicative model's prediction, and four were consonant with the linear and ratio models' prediction. Like the individual-level model fitting results, these qualitative analyses revealed evidence for both the multiplicative and linear models.

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

| | Uninstructed group | | Instructed group | |
|---|---|---|---|---|
| Model | $w$ | VAF | $w$ | VAF |
| Multiplicative | 0.67 | 56.3 | 0.77 | 47.5 |
| Linear | 0.87 | 56.3 | 0.98 | 54.2 |
| Ratio | 0.75 | 52.1 | 0.86 | 47.6 |

*Note.* VAF corresponds to percentage of variance accounted for; $w$ = the degree to which forced-choice confidence was affected by confidence in the nonfocal alternative.

Table 5
*Experiment 2: Percentage of Participants in Each Group for Whom Each Model Accounted for the Most Variance*

| Model | Uninstructed group | Instructed group |
|---|---|---|
| Multiplicative | 44.9 | 37.5 |
| Linear | 46.9 | 47.9 |
| Ratio | 8.2 | 14.6 |

## Discussion

The results of the current experiment, which had participants report confidence in each of the yes–no and forced-choice tasks, are similar in important respects to those of Experiment 1, where confidence judgments were hypothetical. The normative model accounted for less variance than the multiplicative model in both groups, providing evidence that the normative model does not describe change in confidence well. Furthermore, the ratio model's performance was weak across groups, consistent with Experiment 1.

At the group level, the multiplicative model accounted for the most variance for both the uninstructed and instructed groups. At the individual level, however, the linear model did as well as the multiplicative model for the uninstructed group and outperformed the multiplicative model for the instructed group. Thus, as in Experiment 1, the linear model accounted for the most variance at the individual level for the instructed group, with an equal-weighted model performing best. The only difference between the experiments is that the multiplicative model was not the clear winner for the uninstructed group. As we show in the Monte Carlo simulation section, however, the linear model can account for almost as much variance as the multiplicative model, even when the latter is correct. Thus, the strong showing of the linear model may be partly attributable to its inherent flexibility (Myung & Pitt, 1997). Generally, though, the results are consistent across the two experiments in that the multiplicative and linear models appear to be the only viable models, at least among those tested.

Other results are also consistent with the previous experiment: Forced-choice confidence judgments were less additive for the uninstructed group, and this was reflected in the $w$ parameter, which was further from 1 in this group, for each model. That $w$ was less than 1 indicates that the strength of the nonfocal alternative was underweighted. Finally, the actual range of the parameter in the individual-level analyses was again largely as expected.

Bender (1998) reported an experiment intended to generalize a judgment model proposed by Wallsten and González-Vallejo (1994). Doctoral students in history and English literature saw general knowledge statements pertaining to both subject areas. In the first part of the experiment, participants rated their confidence in the truth of individual statements on a 4-point scale. In the second part, the same statements were presented in pairs, with one statement true and one false. In addition, one statement in each pair came from one subject area, and the second statement came from the other. Participants selected the statement they thought was the true one in each pair. Note that the first part is a yes–no task and

the second part is a forced-choice task using the same statements, just like our design. Furthermore, having the paired statements in the forced-choice task come from different domains is also similar to our experiment and underscores the independence of the statements in the yes–no phase. However, Bender's participants did not give confidence ratings in both tasks, which is our focus. Nonetheless, an interesting conclusion from Bender's experiment is that participants did not compare confidence in the two statements when making their choice in the forced-choice task but instead relied on their confidence in the statement from the domain they were more familiar with. If one assumes that the statement from the familiar domain is the focal statement and the statement from the unfamiliar domain is the nonfocal alternative, then Bender's result accords well with one of our consistent findings: Confidence in the nonfocal alternative had less than optimal impact on behavior in the forced-choice task.

## Simulation Analysis of the Role of Error in Judgment

The model fitting results were based on least squares fits of the multiplicative, linear, and ratio models to the forced-choice confidence ratings. Such fits assume that deviations between the observed forced-choice confidence ratings and the predictions of the true model are distributed normally and with equal variance. Because the dependent measure in this case is a proportion, which ranges from 0 to 1, one can be fairly sure that error is not accurately described by a normal distribution, which ranges from $-\infty$ to $+\infty$.[1] One solution to this problem is to transform the independent and dependent measures from probabilities to log odds (which range from $-\infty$ to $+\infty$). This is a natural way to fit the multiplicative model, which, as we demonstrated earlier, can be easily expressed in terms of odds (Equation 4). When we fit the multiplicative model in log odds form, however, the results were very similar to those of the model in its original form. In addition, the linear and ratio models seem rather unnatural when converted to odds form. We therefore decided to leave the data untransformed and to simulate error in a more realistic way. In this section, we describe the results of simulations that investigated whether the (incorrect) assumption of Gaussian error in our least squares fits biased our conclusions in any way.

Having rejected Gaussian error, we considered the important issue of how to model error variance more accurately. We selected a distribution that was appropriately bounded and that mimicked (as described in Footnote 1) our real data. The beta distribution satisfied both requirements. The beta distribution is defined by two parameters, $a$ and $b$, and its range is $0 < p < 1$, the appropriate interval for our purposes. For integer values of $a$ and $b$, the beta distribution is given by $[(a + b)!/(a!b!)](1 - p)^{b-1}(p)^{a-1}$ and has a mean of $a/(a + b)$ and a variance of $ab/[(a + b)^2(a + b + 1)]$. For our simulations, we fixed the sum of $a$ and $b$ at 10 because that generated sufficient error variance such that even the true model accounted for only about 50% of the variance when fit to the simulated data. (This was the approximate percentage of variance accounted for at the individual-participant level in Experiment 2.) Figure 4 shows representative beta distributions for means of 0.5, 0.7, and 0.9 (with the sum of $a$ and $b$ fixed at 10). When the true mean is 0.5, the distribution of values is symmetric and bell-shaped (upper panel). As the true mean approaches 1 (see middle and bottom panels), the variance of the distribution decreases and its shape becomes more skewed. Distributions for means of 0.3 and 0.1 would be mirror reflections of those for 0.7 and 0.9.

An advantage of the simulations is that we were able to program a "true" model to generate hypothetical data with error distributed according to the beta distribution, which is probably a closer approximation to the truth than Gaussian error. We then fit the three models using least squares (with the incorrect assumption of Gaussian error) to the simulated data to see whether the true model accounted for the most variance.

A simulation consisted of the following steps. First, a true model was selected (e.g., the ratio model with $w$ set to 1). Second, 81 pairs of true $c$(A) and $c$(B) values were used to generate $c$(A,B) and $c$(B,A) values from the true model. These 81 pairs consisted of the factorial combination of the numbers 0.1 to 0.9 in steps of 0.1. Third, for a given pair of true $c$(A) and $c$(B) values (e.g., 0.9 and 0.5), actual values to be substituted into the true model were selected from a beta distribution with the appropriate mean. Thus, if the true $c$(A) value was 0.9, a value to be substituted into the true model was drawn from a beta distribution with $a = 9$ and $b = 1$. If the true $c$(B) value was 0.5, a value to be substituted into the true model was drawn from a beta distribution with $a = 5$ and $b = 5$. Thus, although the true $c$(A) and $c$(B) values might be 0.9 and 0.5, the values substituted into the model might be 0.94 and 0.35. Finally, after the true model generated $c$(A,B) and $c$(B,A), these values were replaced by drawing from a beta distribution with the closest mean. Thus, for example, in the ratio model, if $c$(A) and $c$(B) were 0.94 and 0.35, respectively, then $c$(A,B) would equal 0.73. However, the final $c$(A,B) value would not be 0.73 but a value drawn from a beta distribution with a mean of 0.7. In short, we introduced error in confidence reports at both the yes–no and the forced-choice stages.

[1] We examined the assumption of Gaussian error primarily through analyzing quantile–quantile (qq) plots, which provide a check on whether the residuals (i.e., the difference between observed and predicted values) are normally distributed. The two quantiles in such a plot consist of standardized values from the data set (in this case, the residuals) and a corresponding set of values computed from the hypothesized distribution (in this case, the normal distribution). The normal quantiles are calculated by computing the inverse phi function of $(i - .5)/n$, where $i$ is the $i$th largest residual out of $n$ total residuals. To produce a qq plot, the rank-ordered standardized residual values are plotted against these normal quantile values. If the residuals are normally distributed, a straight line plot should be evident. Large and systematic deviations from a straight line indicate that the assumption of Gaussian error is incorrect. Identifying systematic error in a qq plot is something of an art because deviations may appear to be systematic even when the underlying error distribution is Gaussian. Nevertheless, we examined these plots for additional information about the form of underlying error. The qq plots for our individual participants did not deviate noticeably from what would be expected given true Gaussian error (although true error was presumably non-Gaussian). In most cases, the qq plots from Experiment 1 followed a fairly tight-fitting sinusoidal pattern around the best fitting straight line. For some participants, however, the qq plot was a shallow U shape. (Plots for individual participants from Experiment 2 were less informative because participants were not required to use the full range of confidence ratings during the yes–no phase.) Similar kinds of deviation are often observed when qq plots are created based on Gaussian error. The fact that the qq plots from Experiment 1 were not especially deviant suggests that the assumption of Gaussian error in our least squares fits may have been a sufficiently close approximation to avoid arriving at misleading conclusions.

The simulation described above was run 30 times (representing 30 participants) for each of the three models. Thus, for 30 of the simulated participants, the linear model was the true model underlying the data; for another 30, the ratio model was the true model, and for another 30 the multiplicative model was the true model. In every case, all three models were fit to the simulated data using the same method we used to fit the real data.[2]

The top of Table 6 shows the results of the fits for the simulated data when $w = 1$ for each true model (recall that the multiplicative model is normative when $w = 1$), and the bottom shows the results when $w = 0.7$. Five important conclusions emerged from the simulations. First, variance accounted for faithfully reflected the true underlying model in each case. The wrong underlying model never accounted for the most variance. This is why we relied on variance accounted for when discussing which models were and were not ruled out by our experiments.

Second, we mentioned earlier that a model's bias did not help determine which model was correct. In both of our experiments, the linear and ratio models were biased in that participants' confidence was more extreme than predicted by these models. This was true at both the group level (see, e.g., Figure 3) and the individual level of analysis. Though not reported in Table 6, the simulations showed that, when error follows a beta distribution, the linear and ratio models are biased even when they are the correct models. The multiplicative model is the least biased whether or not it is correct. This is why we did not emphasize bias in our earlier discussions.

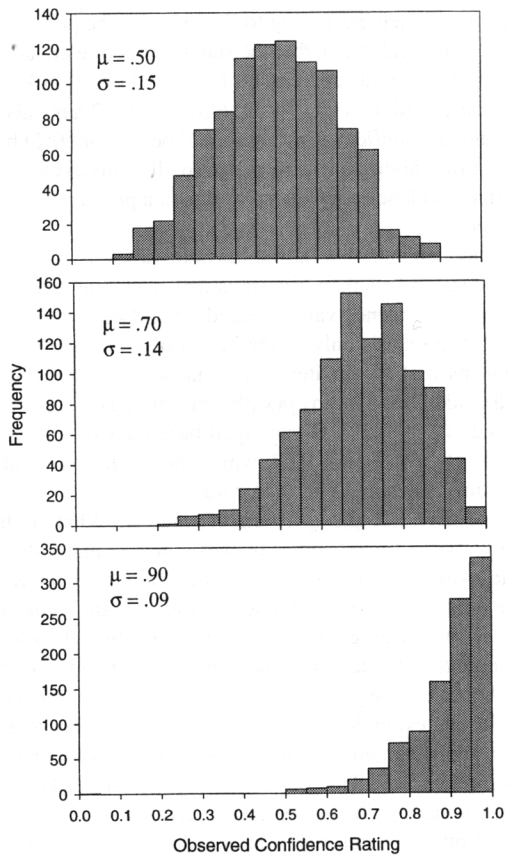


*Figure 4.* Representative beta distributions for means of 0.50, 0.70, and 0.90 (with the sum of the *a* and *b* parameters fixed at 10).

Table 6
*Simulation Results*

| | True (simulated) model with $w = 1.0$ | | | | | |
|---|---|---|---|---|---|---|
| | Multiplicative | | Linear | | Ratio | |
| Fitted model | $w$ | VAF | $w$ | VAF | $w$ | VAF |
| Multiplicative | 0.90 | 50.0 | 0.68 | 29.0 | 0.71 | 29.0 |
| Linear | 0.99 | 45.0 | 1.02 | 54.0 | 1.02 | 48.0 |
| Ratio | 0.95 | 44.0 | 0.93 | 48.0 | 1.00 | 51.0 |
| | True (simulated) model with $w = 0.70$ | | | | | |
| | Multiplicative | | Linear | | Ratio | |
| Fitted model | $w$ | VAF | $w$ | VAF | $w$ | VAF |
| Multiplicative | 0.68 | 52.0 | 0.43 | 42.0 | 0.40 | 39.0 |
| Linear | 0.78 | 48.0 | 0.71 | 54.0 | 0.68 | 50.0 |
| Ratio | 0.77 | 48.0 | 0.68 | 51.0 | 0.67 | 51.0 |

*Note.* VAF refers to percentage of variance accounted for; $w$ = the degree to which forced-choice confidence was affected by confidence in the nonfocal alternative.

Third, Table 6 shows that the models are not equally able to mimic each other. When the true model is the linear or the ratio model, the multiplicative model offers a very poor fit, suggesting that it could be easily ruled out if either the linear or the ratio model were correct. By contrast, when the multiplicative model is the true model, it provides the best fit, but the linear and ratio models are not far behind. This indicates that it might be more difficult to rule out the linear and ratio models based on goodness-of-fit measures alone if the multiplicative model is correct. These results underscore the point that models with the same number of free parameters may differ in their inherent flexibility (Myung & Pitt, 1997).

Fourth, least squares fits assume error-free independent variables ($c$[A] and $c$[B] in the current context). This assumption seems reasonable in Experiment 1 but not in Experiment 2. However, the simulations introduced random error into the independent variables as well as the dependent variables ($c$[A,B] and $c$[B,A]), so the conclusions we drew from Experiment 2 based on the least squares analysis appear valid (again assuming the beta distribution is reasonable). Furthermore, additional simulations (not reported here) were performed using error-free independent variables, and the results were essentially identical. Note that this is consistent with our empirical findings in that the pattern of results is similar in Experiments 1 and 2.

Finally, the top of Table 6 shows that when participants are normative and error variance follows a beta distribution, the estimated value of $w$ from the fit of the multiplicative model is

[2] The qq plots for the simulated data (see Footnote 1) were also inspected to see whether deviations from a straight line were similar to the deviations observed in the real fits. These plots were generally indistinguishable from those produced by our participants and by those based on true Gaussian error (which we also simulated for comparative purposes). Thus, the beta distribution offered a conceptually appealing model of error that was also close enough to Gaussian error so that the least squares fits were able to identify the true underlying model.

somewhat less than 1 (the mean was 0.9 for our 30 simulated participants). However, if the beta distribution offers a reasonable approximation of error, it seems unlikely that error alone can account for the much lower values of $w$ we observed in our experiments. We also simulated 30 additional normative participants without error in the yes–no confidence ratings (analogous to Experiment 1). Otherwise, the simulation was identical to that described above. The mean variance accounted for by fitting the multiplicative model to the simulated data in this case was 83.2% (similar to the actual value observed in Experiment 1), and the mean estimated value of $w$ was 1.04. In other words, error in forced-choice ratings alone does not appear to put any downward pressure on $w$. Because forced-choice ratings were the only source of error in Experiment 1 and the $w$ values in the experiment were nonetheless low, this is additional evidence for rejecting the normative model as a viable descriptive model.

## Signal-Detection Theory and the Normative Model

In the perception and memory literatures, the theoretical framework most commonly used to interpret confidence ratings in yes–no and forced-choice tasks is signal-detection theory. In this section, we consider the relationship between the normative model and the signal-detection interpretation of confidence. The application of signal-detection theory to yes–no and forced-choice tasks has been discussed by Luce (1963) and Ferrell and McGoey (1980). However, the connection between this earlier work and our normative model may not be obvious to most readers. The purpose of this section is to clarify the connection.

Consider first a yes–no task. Typically, half the items on a yes–no test should receive a yes response (e.g., the true statements on a general knowledge test), and half should receive a no response (e.g., the false statements). A signal-detection interpretation of performance assumes that responses are made on the basis of the subjective strength of evidence associated with a test item. The true statements on a general knowledge test, for example, generate a higher subjective value on this scale (i.e., a higher subjective sense that they are true), on average, than the false statements do. Both the true and the false statements are assumed to generate variable strength of evidence values across items. Thus, for example, some true statements do not generate a strong sense that they are true, although most do. Similarly, some false statements generate a strong (though incorrect) sense that they are true, although most do not. The strength of evidence values associated with the two item classes (e.g., true and false statements) can be of any form but are typically assumed to be normally distributed, with the mean of one distribution falling higher on the evidence axis than the mean of the other. This situation is depicted in Figure 5. To solve the task, participants are assumed to set a decision criterion somewhere along the strength of evidence axis such that items falling above the criterion receive a true (or yes) response and items falling below receive a false (or no) response. Note that, according to this model, some true statements (those generating a strength of evidence value falling below the criterion) are mistakenly judged to be false, whereas some false statements (those generating a strength of evidence value falling above the criterion) are mistakenly judged to be true.

What accounts for the varying degrees of confidence across items? According to this model, the higher an item falls on the evidence axis, the more likely it is to be true (and the greater the confidence in a yes response). Consider a hypothetical statement that generates a strength of evidence value that falls at the point labeled $x_A$ in Figure 5. A few false statements generate that level of subjective evidence, but many more true statements do. In fact, at that point, the height of the true distribution is 4 times that of the false distribution. More formally, the likelihood ratio, $p(x_A|\text{A is true})/p(x_A|\text{A is false})$ is equal to 4 (i.e., $L_A = 4$). Thus, if the base rates of true and false statements are equal (as they typically are in an experiment), then, according to Bayes' rule, the posterior odds, $p(\text{A true}|x_A)/p(\text{A false}|x_A)$, that the statement is true are also 4 to 1 (i.e., $O_A = 4$). As indicated earlier, $O_A = c(\text{A})/[1 - c(\text{A})]$, which can be rearranged to $c(\text{A}) = O_A/(O_A + 1)$. Thus, given $x_A$, $O_A = 4$, and the confidence rating should be 0.8 (or 80%) because, in the long run, giving a true response to all items associated with a strength of evidence equal to $x_A$ results in a performance level of 80% correct.

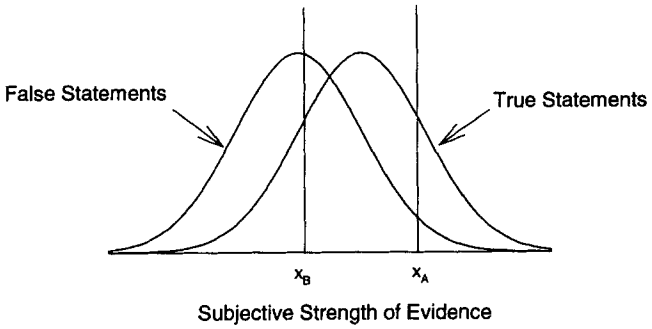


*Figure 5.* A signal-detection view of where two hypothetical statements, A and B, fall on the continuum of subjective strength of evidence. The curve on the left represents the assumed distribution of the subjective strength for false statements, and the curve on the right represents the assumed distribution of the subjective strength for true statements. x corresponds to a hypothetical strength of evidence value.

What about an item that falls at the point labeled $x_B$ in Figure 5? More false statements than true statements generate that relatively low subjective evidence value. Indeed, at that point, the height of the true distribution is only ⅔ the height of the false distribution, which means that for an item with that level of subjective evidence, the likelihood ratio, $p(x_B|\text{B true})/p(x_B|\text{B false})$, is equal to 0.67 (i.e., $L_B = 0.67$). Given equal base rates of true and false statements, $O_B$ also equals 0.67, which means that the confidence rating in this case should be 0.4 or 40%.

Now we may ask the question of interest: What if the task instead involves a forced choice between an item that falls at point A on the evidence axis and an item that falls at point B on the evidence axis? Which item should be selected, and what level of confidence should be expressed? Signal-detection theorists typically assume that the item with the highest likelihood of being true will be chosen (see, e.g., Glanzer et al., 1993; Luce, 1963). Thus, because in our example $L_A > L_B$, item A would be selected as being the true statement. What are the posterior odds that item A is the correct choice? This question is not usually asked, but the answer is implicit in the mathematical derivations offered by Luce (1963) and others. To answer that question, we first compute the likelihoods of obtaining both $x_A$ and $x_B$ under the two possibilities that apply to the forced choice situation: Either A is true and B is false, or A is false and B is true. The likelihood of observing $x_A$

given that A is true is represented by $p(x_A|A \text{ true})$, and the likelihood of observing $x_B$ given that B is false is represented by $p(x_B|B \text{ false})$. Assuming, as we do in this article, that A and B are independent, then the likelihood of observing both $x_A$ and $x_B$ given that A is true and B is false is equal to $p(x_A|A \text{ true})p(x_B|B \text{ false})$. Similarly, the likelihood of observing both $x_A$ and $x_B$ given (alternatively) that A is false and B is true to equal to $p(x_A|A \text{ false})p(x_B|B \text{ true})$. The likelihood ratio for the forced choice case, $L_{A,B}$, is therefore equal to

$$L_{A,B} = \frac{p(x_A|A \text{ true})p(x_B|B \text{ false})}{p(x_A|A \text{ false})p(x_B|B \text{ true})}.$$

Note that the right side of this equation is equal to the ratio of likelihood ratios. That is, $L_{A,B} = L_A/L_B$. The two relevant hypotheses in the forced-choice situation are (a) A is true and B is false, and (b) A is false and B is true. Assuming equal base rates of these two possibilities, the likelihood ratio is equal to the posterior odds (i.e., $L_{A,B} = O_{A,B}$). Thus, $O_{A,B} = O_A/O_B$. Note that this is the normative model derived earlier in odds form (Equation 3). For the example given above in which $O_A = 4$ and $O_B = 2/3$, $O_{A,B}$ is equal to 6. In other words, in the long run, when comparing items that fall at point A and point B on the evidence axis, item A is the correct choice 6 out of 7 times (which corresponds to a confidence rating of 0.86, or 86%). The signal-detection account is equivalent to the normative model (Equation 1) and therefore leads to the same conclusion for our running example: Choosing A with 86% confidence is the normative response.

We have discussed the relationship between our normative model and signal-detection theory, but it should be kept in mind that we rejected the normative model as a descriptive model. This does not mean, however, that we have rejected signal-detection theory as a means of studying our topic, only that participants behave as ideal observers. Indeed, our finding that Equation 3 is not descriptively adequate is consistent with other research. Equation 3 is a special case of Wallsten and González-Vallejo's (1994) stochastic judgment model, which was disconfirmed by Bender (1998) using a task similar to ours (as discussed earlier). Stretch and Wixted (1998) similarly showed that, in a yes–no recognition task, participants adjusted their confidence ratings across various study conditions in a manner that was quantitatively less extreme than, but directionally consistent with, an ideal-observer signal-detection model. A similar conclusion applies here when participants move from a yes–no to a forced-choice task.

## General Discussion

The present research makes four contributions to understanding the relation between confidence in yes–no and forced-choice tasks. First, we derived the normative model for how confidence in two hypotheses should change when they are first presented independently in a yes–no task and then as mutually exclusive and exhaustive competitors in a forced-choice task.

Second, we empirically tested three distinct descriptive models and were able to rule out one, the ratio model. We were not, however, able to rule out the multiplicative and linear models. At the individual level in both experiments, the linear model performed best when participants were instructed to have their forced-choice confidence sum to 100%. When there were no such instructions, the multiplicative model performed best in Experiment 1, and the multiplicative and linear models performed about equally well in Experiment 2. Interestingly, the additivity instructions appeared to lead participants to use not just a linear strategy but an equal-weighted linear strategy. Though linear models are known for being able to mimic nonlinear processes well (which our computer simulations also showed), qualitative analyses corroborated the model fitting results, indicating that it is not just the robustness of the linear model that accounts for its success. Also of interest is that the free parameter in the multiplicative model was consistently less than 1, implying that confidence in the alternative statement had less than optimal impact when reporting confidence in the forced-choice task. Importantly, this latter finding also allowed us to reject the normative model as a descriptive model.

Third, our least squares fits assumed Gaussian error in participants' confidence reports, but computer simulations using a more realistic representation of error (the beta distribution) supported our conclusions from Experiments 1 and 2. The simulations were critical for determining which measures of performance were and were not helpful for deciding which models could be ruled out. More generally, the simulations highlighted the importance of understanding the implications of random error in confidence reports (Erev, Wallsten, & Budescu, 1994; Soll, 1996).

Finally, we demonstrated the equivalence of the normative model and the signal-detection analysis of confidence. We also recast the normative model and the multiplicative model in terms more familiar to signal-detection theorists (likelihoods and odds; see Equations 3 and 4) to facilitate use of the models in that area.

A consistent empirical finding is that confidence judgments for hypotheses in the forced-choice tasks were nonadditive in the absence of explicit instructions to be additive. Many readers may find this surprising, given that it was made clear that statements forming each pair were mutually exclusive and exhaustive. Indeed, the assumption of additivity on the part of researchers is so strong that they rarely ask for confidence estimates for all items in a forced-choice task. Furthermore, there have been several empirical demonstrations of additivity for two mutually exclusive and exhaustive hypotheses (see, e.g., Wallsten et al., 1993), and such additivity is a fundamental implication of a recent influential theory of subjective probability (Tversky & Koehler, 1994). Recall, though, that we measured additivity in terms of mean absolute deviation between summed confidence and 100 across all pairs, whereas additivity is typically measured in terms of mean summed confidence. As discussed earlier, the latter measure can mask dramatic nonadditive confidence estimates, and we feel that the former measure is more appropriate, at least under the current circumstances.

As for why the nonadditivity occurred, the most natural interpretation of the free parameter in the multiplicative model is that it reflects the extent to which the strength of the nonfocal alternative is taken into account. The parameter consistently indicated that the nonfocal alternative was underweighted, and one result of this is nonadditivity (McKenzie, 1998, 1999). Another possible reason for the nonadditivity, however, is that participants did not fully believe the stipulation that the two options at the forced-choice stage were mutually exclusive and exhaustive. Thus, one avenue of future inquiry is to determine whether participants believe the stipulation but discount the nonfocal alternative anyway (as the multiplicative model assumes) or whether they do not fully believe it and report their confidence in the options accordingly.

Future research that can distinguish better between the linear and the multiplicative models will also be useful. Our results indicate that additivity instructions led participants to use an equal-weighted linear rule for reporting confidence in forced-choice tasks. Qualitative analyses also favored the linear model, providing evidence that the linear model's good quantitative performance is not due entirely to its ability (shown by the simulations) to mimic the multiplicative model. Nonetheless, because the linear model does mimic the multiplicative model so well, experiments making more use of the critical tests we discussed earlier would probably be best.

Despite the wide variety of contexts in which yes–no and forced-choice tasks are used, we are reasonably confident that the linear and multiplicative models will outperform the ratio and normative models (though we cannot rule out the possibility that an entirely different model from those we have examined is the best descriptive model). Our results are reasonably consistent across two experiments that, though dealing with similar content, were different in that one dealt with hypothetical statements and confidence and the other with actual statements and confidence. Nonetheless, different content areas undoubtedly affect the models' overall performance. Despite our relatively similar content and context, we found that the models performed somewhat differently between our two experiments, as well as between groups within the same experiment. A perception experiment regarding the presence of visual stimuli might result in different parameter values and amount of variance accounted for than our experiments using general knowledge statements. The main factor that affects the models' free parameter appears to be the extent to which participants take into account the strength of the nonfocal alternative. Factors that affect variance accounted for include the amount of random error in judgment and the extent to which the entire response scale is used. How the models perform across the variety of domains in which yes–no and forced-choice tasks are used is an open empirical question.

We stated at the beginning of this article that little is known about how confidence in yes–no and forced-choice tasks is related. We believe that we have made progress but have only scratched the surface. Considerable theoretical and empirical work remains. Our hope is that we have provided a promising starting point and perhaps a nudge in the right direction.

## References


Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

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

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

Creelman, C. D., & Macmillan, N. A. (1979). Auditory phase and frequency discrimination: A comparison of nine procedures. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 146–156.

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

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95–106.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13,* 171–192.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519–527.

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

Ferrell, W. R., & McGoey, P. J. (1989). 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.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564.

Fletcher, R. (1972). *FORTRAN subroutines for minimization by quasi-Newton methods.* Harwell, England: United Kingdom Atomic Energy Authority.

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

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition 16,* 5–16.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567.

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.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition, 24,* 608–628.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18,* 441–461.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65,* 137–165.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57,* 116–131.

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

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57,* 226–246.

Kim, K., & Glanzer, M. (1993). Speed versus accuracy instructions, study time, and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 638–652.

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

Koehler, D. J. (1996). A strength model of probability judgments for tournaments. *Organizational Behavior and Human Decision Processes, 66,* 16–21.

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

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter, (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103–189). New York: Wiley.

McCloskey, M., & Glucksberg, S. (1978). Natural categories; Well defined or fuzzy sets? *Memory & Cognition, 6,* 462–472.

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.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review, 4,* 79–95.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785.

Robinson, L. B., & Hastie, R. (1985). Revision of beliefs when a hypothesis is eliminated from consideration. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 443–456.

Roediger, H. L. III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes, 65,* 117–137.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1397–1410.

Swets, J. A., Markowitz, J., & Franzen, O. (1969). Vibrotactile signal detection. *Perception and Psychophysics, 6,* 83–88.

Teigen, K. H. (1983). Studies in subjective probability: III. The unimportance of alternatives. *Scandinavian Journal of Psychology, 24,* 97–105.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547–567.

Van Wallendael, L. R. (1989). The quest for limits on noncomplementarity in opinion revision. *Organizational Behavior and Human Decision Processes, 43,* 385–405.

Van Wallendael, L. R., & Hastie, R. (1990). Tracing the steps of Sherlock Holmes: Cognitive representations of hypothesis testing. *Memory & Cognition, 18,* 240–250.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probabilistic judgments. *Management Science, 39,* 176–190.

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review, 101,* 490–504.

## Appendix

## Derivation of the Normative Model

In this appendix, we derive the normative model for combining confidence estimates given the new knowledge that two propositions are mutually exclusive and exhaustive. This model initially appears in Equation 1. We show here that this formula is Bayes optimal under a few basic assumptions.

Let A be the event that the first of a given pair of statements is true, and let B be the independent event that the second statement in the pair is true. Let X be the event that A and B are mutually exclusive and exhaustive—that exactly one of the two statements 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 $p(\cdot)$ be an appropriate probability mass function over these events.

We want to compute the value of $c(A,B)$, which can normatively be viewed as the probability of A given the fact that A and B are mutually exclusive and exhaustive. In Bayesian terms, this is the value of $p(A|X)$, which may be calculated as follows:

$$p(A|X) = p(A \wedge X)/p(X)$$

$$= p(A \wedge \bar{B})/p([A \wedge \bar{B}] \vee [\bar{A} \wedge B])$$

$$= \frac{p(A \wedge \bar{B})}{p(A \wedge \bar{B}) + p(\bar{A} \wedge B)}$$

$$= \frac{p(A|\bar{B})p(\bar{B})}{p(A|\bar{B})p(\bar{B}) + p(\bar{A}|B)p(B)}.$$

This is the normative model if one does not assume that A and B are independent (prior to knowledge concerning the truth of X). When the independence assumption is made (as in this article), the expression may be simplified to:

$$p(A|X) = \frac{p(A)p(\bar{B})}{p(A)p(\bar{B}) + p(B)p(\bar{A})}$$

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

Because $p(A|X)$ is the normative correlate of $c(A,B)$ and $p(A)$ and $p(B)$ are the normative correlates of $c(A)$ and $c(B)$, respectively, this derivation shows that Equation 1 is normative in a Bayesian sense.

Received February 9, 1999
Revision received August 4, 1999
Accepted October 15, 1999 ■