ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 71, No. 2, August, pp. 141–160, 1997
ARTICLE NO. OB972716

Underweighting Alternatives and Overconfidence

Craig R. M. McKenzie

University of California at San Diego

Much evidence outside the overconfidence literature indicates that confidence in a particular hypothesis is influenced more by evidence for and against that hypothesis than by evidence for and against the alternative. This article focuses on the impact that underweighting the alternative has on overconfidence and other related measures. Computer simulations showed that underweighting the alternative is sufficient for producing overconfidence under quite general conditions. In addition, data from two previous empirical studies were reanalyzed. In these simulated medical diagnosis studies, one group of subjects was known to take into account the alternative, and one to underweight it. The pattern of differences between the two groups was similar to the pattern found in the computer simulations where weighting of the alternative was manipulated. Furthermore, encouragement to take into account the alternative had no effect on the former group's behavior, but affected the latter group in a manner predicted by the simulations. Both the simulations and reanalyses provide direct evidence of the effect of underweighting the strength of alternatives on overconfidence.

In a typical task investigating over/underconfidence, subjects are asked questions and presented with two potential answers, exactly one of which is true. An example is "Absinthe is (a) a precious stone, or (b) a liqueur." In some experiments, subjects select the answer they think is correct, then report confidence on a scale of 50 to 100% (or .5 to 1.0). In other experiments, one answer is preselected and subjects report confidence in its truth on a scale of 0 to

This research was supported by NSF Grant SBR-9515030. I thank two anonymous reviewers for helpful comments and John Wixted for help with Equation 4.

Address correspondence and reprint requests to the author at the Department of Psychology, University of California at San Diego, La Jolla CA 92093-0109. E-mail: cmckenzie@ucsd.edu.

0749-5978/97 $25.00

100%. The type of task also varies. The above “absinthe” question is an example of a general knowledge task, perhaps the most common type in the overconfidence literature. Other tasks include predicting the outcome of events (e.g., winners of basketball games; Ronis & Yates, 1987) and diagnosing causes of events (e.g., physicians diagnosing pneumonia; Christensen-Szalanski & Bushyhead, 1981).

The primary motivation behind such experiments is to investigate subjects’ calibration. One is well calibrated if correct X% of the time when reporting X% confidence. A common finding is that people are not well calibrated; specifically, people tend to be overconfident. In half-scale tasks, this is evidenced by confidence that is too high relative to percent correct. For example, when reporting 90% confidence, subjects are typically correct between 70 and 80% of the time. In full-scale tasks, overconfidence is manifest in confidence too high when greater than 50%, and too low when less than 50%.

Several explanations of overconfidence have been proposed. Some authors have argued that general knowledge questions are implicitly selected for difficulty, thereby leading to overconfidence (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994). Other authors have recently pointed out that random noise in subjects’ judgments can, by itself, lead to overconfidence (Erev, Wallsten, & Budescu, 1994; Soll, 1996). Still others have used signal detection theory to argue that the general finding of overconfidence is due to subjects’ inability to adjust their response criteria in the absence of feedback (e.g., Ferrell & McGoey, 1980).

This article focuses on yet another explanation that has been frequently invoked, namely, that subjects, when reporting confidence in a particular answer, tend to overweight evidence regarding that answer. (This is not to deny the contribution of the above accounts; these explanations are not mutually exclusive—nor exhaustive.) For example, if one selected the “liqueur” answer (or it was preselected), one might consider why that answer might be correct and pay relatively little attention to why the “precious stone” answer might be correct. The first section examines this notion in some detail and points out different ways that such biased processing could occur. The current emphasis is on one particular process, namely, considering evidence both for and against the focal answer, while ignoring or underweighting evidence for and against the alternative. The second part shows through computer simulations that such a process results in overconfidence under some general conditions. The third section discusses two ways of cognitively representing confidence, one of which distinguishes between evidence for one answer and evidence against the other, and one of which does not (McKenzie, in press). The different representations have implications for the extent to which the strength of the alternative is taken into account, which implies that the different representations for the same set of events should lead to different degrees of overconfidence. Data from two experiments (McKenzie, in press) are reanalyzed in order to test these predictions. The final section discusses implications of the findings.

CONSIDERATION OF THE ALTERNATIVE

The task of current interest involves two mutually exclusive and exhaustive hypotheses, X and Y. Confidence tends to be reported in just one hypothesis, which will be referred to as the focal hypothesis. The remaining hypothesis is the alternative. Consider first each hypothesis as a separate, independent statement; that is, consider when the two hypotheses are neither mutually exclusive nor exhaustive. Using the absinthe example, the separate statements would be "Absinthe is a precious stone" (X) and "Absinthe is a liqueur" (Y). Though we will ultimately be concerned with tasks in which X and Y are competing hypotheses, subjects could be presented with each statement separately and report confidence in the truth of each. From the subjects' perspective, X and Y could both be true or both false under such circumstances and therefore one could conceivably have either high or low confidence in each. In other words, one would not necessarily expect confidence in the two hypotheses to be additive, or sum to 100%.

It is assumed that a hypothesis (or statement) gives rise to covert confidence, which is then translated into an overt response (e.g., Erev *et al.*, 1994; Koriat, Lichtenstein, & Fischhoff, 1980). Covert confidence, $s(X)$, or strength of X, is assumed to be an increasing function of evidence for X and a decreasing function of evidence against X. That is,

$$s(X) = f(\text{evidence for } X, \text{ evidence against } X), \tag{1}$$

where f is defined as increasing in its first argument and decreasing in its second argument, and $0 < s(X) < 1$. Similarly, $s(Y) = f(\text{evidence for } Y, \text{ evidence against } Y)$, where $0 < s(Y) < 1$. Furthermore, it is assumed that the overt response, $\text{conf}(X)$, is an increasing function of $s(X)$. That is, $\text{conf}(X) = g[s(X)]$, where g is an increasing function, and $0 < \text{conf}(X) < 1$. For simplicity, it will be assumed that $\text{conf}(X) = s(X)$ (but see Erev *et al.*, 1994; Soll, 1996).

Now consider the case where X and Y are competing, mutually exclusive and exhaustive hypotheses. Confidence in one hypothesis should be an increasing function of the strength of that hypothesis and a decreasing function of the strength of the alternative; that is, normatively speaking,

$$\text{conf}(X,Y) = f[\text{s}(X), s(Y)] = f[\text{conf}(X), \text{conf}(Y)], \tag{2}$$

where $\text{conf}(X,Y)$ refers to confidence in X when Y is the alternative. Assuming that $\text{conf}(X,Y)$ and $\text{conf}(Y,X)$ should (a) be additive (i.e., $\text{conf}(X,Y) + \text{conf}(Y,X) = 1$) and (b) maintain the same ratio as the separate confidence judgments (i.e., $\text{conf}(X,Y)/\text{conf}(Y,X) = \text{conf}(X)/\text{conf}(Y)$), then $\text{conf}(X)$ and $\text{conf}(Y)$ should be normalized to arrive at a normative response:

$$\text{Conf}(X,Y) = \text{conf}(X)/[\text{conf}(X) + \text{conf}(Y)]. \tag{3}$$

The uppercase c is used to denote that the output is normative, given the above assumptions. Note that as conf(Y) increases, Conf(X,Y) decreases.

Figure 1 shows evidence for and against each hypothesis in 2×2 form. Cells A and B represent, respectively, evidence for and against X, and Cells C and D represent, respectively, evidence for and against Y. Subjective confidence, conf(X,Y), should be an increasing function of Cells A and D, and a decreasing function of Cells B and C. Furthermore, each cell should be equally relevant. For example, Cell C evidence should not be discounted relative to Cell A evidence.

What does it mean to overweight evidence regarding the focal hypothesis? It can mean at least three things. Koriat *et al.* (1980) had subjects generate reasons in each of the four cells in Fig. 1 before making a choice and reporting confidence. Subjects also rated each reason in terms of strength. Several analyses revealed a similar pattern: Cell A reasons had the largest impact on confidence (assuming X is chosen, or focal), Cell D second largest, and Cells B and C had almost no impact. Here, then, overconfidence was seen as resulting from overweighting evidence supporting the focal hypothesis (i.e., A&D $>$ B&C).

Another possibility is that subjects tend to consider only evidence for each alternative, Cells A and C, but that A (when X is focal) tends to have a bigger impact on confidence than C (Sniezek, Paese, & Switzer, 1990, pp. 273–274, appear to imply such a process). This, too, could lead to overconfidence.

There is still another possibility. Subjects might consider only evidence for and against the focal hypothesis (i.e., Cells A and B), or at least overweight it relative to evidence for and against the alternative (C and D). Some authors have suggested that such a process might be responsible, at least in part, for overconfidence (Arkes, Christensen, Lai, & Blumer, 1987; Ronis and Yates, 1987; see also Koehler, 1994). Though there is little direct evidence to support this, there is considerable evidence outside the overconfidence literature consistent with this process. For example, a large number of studies have examined how people assess covariation between two dichotomous variables. Such tasks can be represented in 2×2 form like Fig. 1 and can often be thought of in terms of testing two hypotheses, X and Y, with X focal (McKenzie, 1994). A common finding is that Cells A and B have the largest impact on judgments. These conclusions have been reached in studies that have (a) examined subjects' ratings of the four cells in terms of importance (Crocker, 1982; Wasserman *et al.*, 1990), (b) regressed subjects' judgments of covariation onto cell frequencies (Schustack & Sternberg, 1981; see also Lipe, 1990), and (c) inferred which

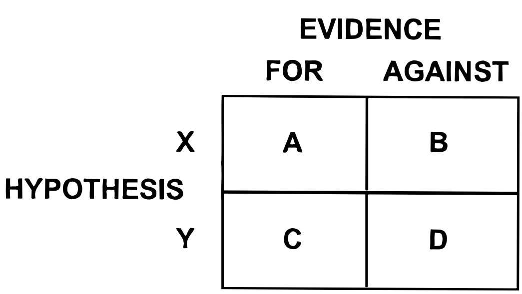

FIG. 1. Evidence for and against two hypotheses.

"rules" subjects appeared to use based on their patterns of responses (Ward & Jenkins, 1965; Arkes & Harkness, 1983; Wasserman *et al.*, 1990). More generally, Klayman and Ha (1987) have argued that hypothesis testing behavior can be interpreted as reflecting a tendency for (in this context) *positive hypothesis testing.* This essentially means testing cases expected to work under the focal hypothesis and see if they do; cases not expected to work under the alternative tend not to get tested. Thus, evidence for and against the focal hypothesis (Cells A and B) is emphasized relative to evidence for and against the alternative (C and D). Along similar lines, subjects are also influenced by pseudodiagnostic information (Doherty *et al.*, 1979). When selecting likelihood information in a Bayesian task in order to determine if a focal hypothesis (X) is true, subjects tend to opt for multiple likelihoods regarding X and ignore likelihoods regarding Y (see also Beyth-Marom & Fischhoff, 1983). This behavior can also be interpreted as a preference for Cells A and B over C and D (McKenzie, 1994).

Thus, there are at least three ways that subjects might overweight evidence regarding the focal hypothesis: A&D > B&C; A > C; or A&B > C&D. It seems clear that overconfidence could result from the first two because evidence supporting the focal hypothesis has a larger impact than evidence supporting the alternative. It is not so clear, though, that the last process—which will be referred to as underweighting the strength of the alternative—will result in overconfidence because though evidence for the alternative (C) is underweighted, so is evidence against the alternative (D). The net effect, in general, might be zero.

Assume that $\text{conf}(X,Y) = \text{conf}(X)$; that is, the strength of the alternative is ignored. Assume further that confidence in the separate X and Y hypotheses is well calibrated. If $\text{conf}(X) = .7$, overconfidence results only if $\text{conf}(Y) > .3$. If $\text{conf}(Y) < .3$, then *under*confidence would result. Thus, given that the ignored alternative could be either stronger or weaker than the subject's response "assumes," overconfidence is not a necessary result. But it is a likely result. The reason is that, given some reasonable assumptions, the ignored alternative will, *on average*, be stronger than the subject's response "assumes." If $\text{conf}(X,Y) = \text{conf}(X) = .7$, though $\text{conf}(Y)$ might sometimes be less than .3, it will, on average, be greater than .3, leading to overconfidence. Generally speaking, when confidence in the focal hypothesis is above .5, the alternative will not be weak enough to justify such high confidence. Similarly, when confidence is less than .5, the alternative will not be strong enough, on average, to justify such low confidence.

COMPUTER SIMULATIONS

It is perhaps easiest to demonstrate the above conclusion through computer simulation. Assume two mutually exclusive and exhaustive hypotheses, X and Y. Let $\text{conf}(X)$ and $\text{conf}(Y)$ vary independently and uniformly between .01 and .99 (resulting in 99^2, or 9801, observations) and let Eq. (3) represent the normative response. It is assumed that confidence in the separate hypotheses evaluated independently is perfectly calibrated (i.e., miscalibration in the separate hypotheses is not necessary for the conclusions reached below). Following is the model used in the simulations to represent subjective confidence:

$$\text{conf}(X,Y) = \text{conf}(X)/[\text{conf}(X) + \text{conf}(Y)]^w, \quad (4)$$

where $0 \leq w \leq 1$. When $w = 0$, $\text{conf}(X,Y) = \text{conf}(X)$; that is, the strength of the alternative has no impact on confidence in the focal hypothesis. When $w = 1$, confidence is normalized and equals the normative response. For example, if $\text{conf}(X) = \text{conf}(Y) = .8$, then $\text{conf}(X,Y) = \text{conf}(Y,X) = .5$ for $w = 1$. For any $w > 0$, stronger alternatives lead to lower confidence in the focal hypothesis. Furthermore, the larger w is, the larger the impact the alternative has on confidence in the focal hypothesis. Values of w between 0 and 1 can be interpreted as various degrees of taking into account the alternative.

In empirical studies examining when and how alternatives are taken into account (summarized below), considering the alternative affects confidence in a manner implied by Equation 4; that is, confidence appears to become normalized. For example, if confidence in each hypothesis is otherwise high ($> .5$), encouragement to take into account the alternative leads to lower confidence in each. If confidence is low in each hypothesis ($< .5$), considering the alternative leads to higher confidence (McKenzie, in press).

Computer simulations were conducted with different values of w, and the results for when $w = 1$, .5, and 0 are shown in Fig. 2. Along the x axis is subjective confidence (Eq. (4)). For simplicity, if $\text{conf}(X,Y)$ was less than .5, it was transformed to $1 - \text{conf}(X,Y)$. Confidence was then grouped into five categories: .55, .65, . . ., .95. The y axis represents the normative response (Eq.

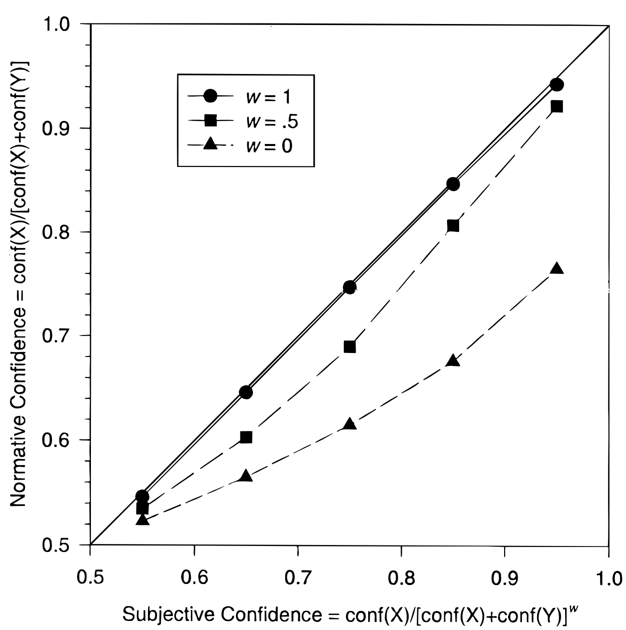

FIG. 2. Calibration curves from the computer simulations for $w = 1$, .5, and 0.

(3)). Plotted is the mean normative response for each confidence category. Perfect calibration entails that each point lie on the identity line. As one would expect, calibration was perfect when $w = 1$. (Departures from the identity line are due solely to the fact that category labels—e.g., ".55"—do not correspond exactly to mean confidence for the category.) When $w = .5$, the alternative was underweighted and all the points lie below the identity line, indicating overconfidence. When $w = 0$, the strength of the alternative was ignored and even greater overconfidence resulted.

Figure 2 represents a traditional way of reporting overconfidence results, where confidence is grouped. When objective probabilities (i.e., normative responses) are known, as in the computer simulations (and in the experiments to be reanalyzed), grouping is not necessary. The top of Table 1 reports various measures (defined in Table 2) from the computer simulations using ungrouped confidence. The first measure is mean confidence (on a scale of .50 to .99), followed by expected proportion correct, E(correct), which is the mean normative response for the hypotheses supported by the subjective confidence. (If conf(X,Y) was $\geq$ .5, X was supported, else Y was supported.) The next measure is overconfidence, which equals the signed difference between the first two measures. The fourth measure is the mean square deviation (MSD) between subjective confidence and the normative response across all observations and is a general measure of accuracy. Finally, there is a measure of additivity: the mean absolute deviation (MAD) between 1 and the sum of conf(X,Y) and conf(Y,X) across all X,Y pairs.

The pattern of results as w decreased from 1 to 0 is clear: conf(X,Y) increased and E(correct) decreased, leading to increased overconfidence. Furthermore, both MSD and MAD increased, indicating poorer accuracy and decreased additivity, respectively.

TABLE 1

Performance Measures for Computer Simulations and Reanalysis of McKenzie (in press), Experiments 3 and 4

	Confidence	*E*(Correct)	Overconfidence	*MSD*	*MAD* from 1
		Computer simulations			
$w = 1.0$	.690	.690	.000	.000	.000
$w = 0.5$	.700	.664	.037	.015	.170
$w = 0.0$	.747	.630	.118	.045	.330
		McKenzie (in press), Experiment 3			
CL	.790	.665*	.125*	.092*	.167*
NCL	.773	.599	.175	.132	.390
		McKenzie (in press), Experiment 4			
CL					
Sym Qs	.803	.674	.128	.091	.163
Asym Qs	.816	.659	.157	.101	.182
NCL					
Sym Qs	.732*	.594	.138*	.127*	.268*
Asym Qs	.772	.563	.209	.156	.422

Note. Values with an asterisk are different from values immediately below them, $p < .05$.

TABLE 2

Definitions of Dependent Measures

Mean confidence	
	$1/N \sum \text{conf}(X_i, Y_i)$
E(correct)	
	$1/N \sum \text{Conf}(\cdot_i, \cdot_i)$, where
	if $\text{conf}(X_i, Y_i) \geq .5$, then $\text{Conf}(\cdot_i, \cdot_i) = \text{Conf}(X_i, Y_i)$
	if $\text{conf}(X_i, Y_i) < .5$, then $\text{Conf}(\cdot_i, \cdot_i) = \text{Conf}(Y_i, X_i)$
Overconfidence	
	Mean confidence $-$ E(correct)
Mean square deviation (MSD)	
	$1/N \sum [\text{conf}(X_i, Y_i) - \text{Conf}(X_i, Y_i)]^2$
Mean absolute deviation (MAD)	
	$1/N \sum \lvert 1 - [\text{conf}(X_i, Y_i) + \text{conf}(Y_i, X_i)] \rvert$

Note. Conf(·,·) (uppercase c) corresponds to normative confidence (Eq. (3) for the simulations, Table 6 for the experiments) and conf(·,·) (lowercase c) to subjective confidence (Eq. (4) for the simulations). $N = 9801$ for the simulations and 32 for the experiments. Mean confidence is based on half-scale $\text{conf}(X_i, Y_i)$: If $\text{conf}(X_i, Y_i) < .5$, then its complement was used.

In order to check the robustness of the findings, other simulations were conducted with various nonuniform distributions. In one series, the distributions were U-shaped (ten .01 and ten .99 probabilities were added), corresponding to situations where reported confidence and true probabilities tend to be extreme. As w decreased, conf(X,Y) increased and E(correct) decreased, with the latter change more extreme than when the distributions were uniform, resulting in even greater overconfidence when $w < 1$. Another series used W-shaped distributions (in addition to the above, ten .5 probabilities were added), corresponding to confidence and true probabilities “bunching up” at the endpoints and middle. The pattern of change in conf(X,Y) and E(correct) as w decreased was again the same, with resulting overconfidence falling between that of the uniform and U-shaped distributions. A final series used inverted-U-shaped distributions (only ten .5 probabilities were added) and resulted in slightly less overconfidence than the uniform distributions. If the strength of the alternative is underweighted or ignored, extreme probabilities result in more overconfidence, while adding moderate ones decreases overconfidence. Nonetheless, the pattern was the same across all four distributions: As w decreased, conf(X,Y) increased and E(correct) decreased, resulting in more overconfidence. Furthermore, MSD and MAD increased as w decreased in every case.

The assumption that conf(X) and conf(Y) are independent was also examined. One series of simulations used uniform distributions with a correlation of $-.22$. A negative correlation between conf(X) and conf(Y) might be reasonable to expect, even when X and Y are evaluated independently and (from the hypothetical subjects’ viewpoint) are neither mutually exclusive nor exhaustive. Nonetheless, the same pattern emerged, albeit slightly weaker relative to the independent distributions: As w decreased, conf(X,Y) increased and E(correct) decreased, leading to more overconfidence; MAD and MSD also increased.

Finally, when conf(X) and conf(Y) were positively correlated, overconfidence was greater relative to the independent distributions for $w < 1$.

Thus, not only do independent distributions symmetric around .5 appear sufficient for these results, independence is not necessary. The findings are important because they illustrate that when the strength of the alternative hypothesis is ignored or underweighted, overconfidence is the result under quite general conditions.

DEPENDENT VERSUS INDEPENDENT CONFIDENCE

Confidence in two hypotheses can be cognitively represented in one of two ways (McKenzie, in press; see also Van Wallendael & Hastie, 1990). First, confidence can be dependent, illustrated in Fig. 3a. The poles of the dependent scale correspond to the truth of each hypothesis. Any change in confidence in one hypothesis necessitates a complementary change in the other; for example, an increase in confidence that *X* is true decreases confidence in the truth of *Y*. Second, confidence can be independent (Fig. 3b), where each hypothesis can be thought of as having a separate scale. The poles of each scale correspond to the truth and falsity of the respective hypotheses. Confidence can increase or decrease in one hypothesis, but not change in the other. Indeed, confidence can, in theory, increase or decrease in both hypotheses simultaneously.

The distinction between the two representations has implications for when (and how) the strength of the alternative is taken into account. With dependent confidence, taking into account the alternative is part of the nature of the process because any change in confidence regarding one hypothesis directly affects confidence in the other. For example, an increase in confidence in *X* and an equally strong increase in *Y* would simply offset each other. With independent confidence, however, such changes might lead to an increase in confidence on each separate scale. Taking into account the strength of the alternative would require accessing—and somehow accommodating—confidence in the alternative as well as the focal hypothesis. Ignoring the

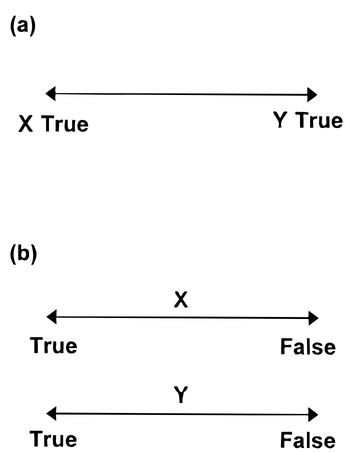

FIG. 3. (a) Dependent and (b) independent confidence in two hypotheses.

alternative (i.e., simply "reading off" the focal hypothesis scale) in this case would result in reporting confidence too high in whichever hypothesis was focal. Thus, consideration of the alternative requires additional cognitive processing when confidence is independent. Because the extra processing might not take place, taking into account the alternative is less likely with independent confidence.

These different representations arise from different conceptions of evidence (McKenzie, in press). Dependent confidence is the result of conceiving of just two kinds of relevant evidence: that for *X* and that for *Y* (Fig. 4a). The former leads to movement to the left on the dependent scale, the latter leads to movements to the right. There is, psychologically speaking, no distinction between evidence for one hypothesis and evidence against the other. For example, when assessing conf(*X*,*Y*), evidence for *X* and evidence against *Y* have (roughly) the same impact (McKenzie, 1997, in press). Thus, $\text{conf}(X,Y) = f[s(X), s(Y)] = f[\text{conf}(X), \text{conf}(Y)]$; similarly, $\text{conf}(Y,X) = f[s(Y), s(X)] = f[\text{conf}(Y), \text{conf}(X)]$. The same evidence is deemed relevant regardless which hypothesis is focal. In contrast, independent confidence results from conceiving of four kinds of evidence (Fig. 4b), where there is a psychological distinction between evidence for one hypothesis and evidence against the other. Evidence for (against) *Y* is not necessarily seen as evidence against (for) *X*, and conf(*X*,*Y*) can equal conf(*X*). Put differently, conf(*X*,*Y*) can be a function of evidence for and against *X* without regard to evidence for and against *Y*. Different evidence might be evoked when evaluating conf(*X*,*Y*) and when evaluating conf(*Y*,*X*).

Because taking into account the strength of the alternative is less likely with independent confidence, such a representation should lead to more overconfidence than a dependent representation of confidence in the same hypotheses. Data from two experiments (McKenzie, in press) are reanalyzed in order to test these predictions.

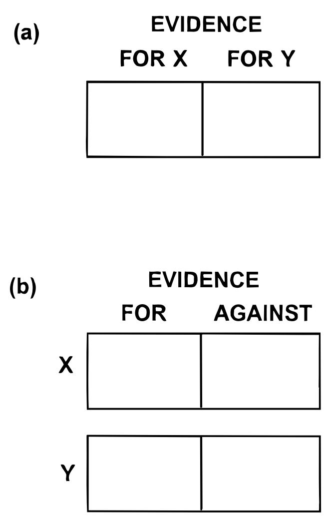

FIG. 4. (a) Two- and (b) four-category conceptions of evidence in two hypotheses.

TABLE 3

Example of a Patient Profile

Symptom	Present?
Coughing	No
Rash	Yes
Nausea	Yes
Fever	No

OVERVIEW OF EXPERIMENTS TO BE REANALYZED

How can different representations of confidence in the same hypotheses be produced? A learning manipulation has been used previously (McKenzie, 1997, in press; see also Goldstone, 1996; Klayman & Brown, 1993). In McKenzie (1997, in press, Experiments 3 and 4), subjects learned about two fictitious illnesses, puneria and zymosis, through viewing patient profiles listing the presence and absence of four symptoms (an example is shown in Table 3). The *contrastive* group learned to distinguish between the illnesses. Contrastive learning (CL) is illustrated in Table 4. On the left side are four symptoms, S1 through S4. The other two columns list the proportion of patients who exhibit the symptoms, given each illness. This side-by-side contrasting makes clear the diagnosticity of each symptom for the two illnesses. For example, it is obvious that S1 is simultaneously strong evidence for puneria and against zymosis because it is common in the former and rare in the latter. Similarly, it is clear that S2 is both weak evidence for puneria and against zymosis, and so on. CL was instantiated through having subjects view 60 puneria and 60 zymosis profiles in a random order, make a diagnosis after each, and then receive feedback. In this way, CL subjects learned to distinguish between the illnesses.

The *noncontrastive* group learned about the illnesses separately. The top of Table 5 illustrates learning about puneria, where patients with and without the illness are contrasted. Patients without puneria tend to have none of the symptoms and can be thought of as healthy. S1 through S4 are learned to be decreasingly diagnostic of puneria, with S1 strong evidence for puneria and S4 nondiagnostic. The bottom of Table 5 shows learning about zymosis, where patients with and without zymosis are contrasted. Here, S1 through S4 are

TABLE 4
Illustration of Contrastive Learning of Puneria and Zymosis

Symptom	Puneria	Zymosis
S1	.80	.20
S2	.60	.40
S3	.40	.60
S4	.20	.80

TABLE 5
Illustration of Noncontrastive Learning of Puneria and Zymosis

Puneria		
Symptom	Puneria	No Puneria
S1	.80	.20
S2	.60	.20
S3	.40	.20
S4	.20	.20

Zymosis		
Symptom	Zymosis	No Zymosis
S1	.20	.20
S2	.40	.20
S3	.60	.20
S4	.80	.20

learned to be increasingly diagnostic of zymosis, with S1 nondiagnostic and S4 strong evidence for zymosis. When learning about puneria, noncontrastive learning (NCL) subjects saw 60 puneria and 60 "healthy" profiles, made a diagnosis, then received feedback. They learned about zymosis analogously. Both CL and NCL subjects saw 60 profiles of each illness, performed their respective tasks equally well, learned across trials, reached asymptote before the end of the trials, and rated learning equally difficult.

CL and NCL lead to different qualitative understandings of the relation between the symptoms and the illnesses. CL leads to learning that S1 and S2 are evidence for puneria and against zymosis, whereas NCL leads to learning that S1 is evidence for puneria but nondiagnostic with regard to zymosis, and that S2 is evidence for both illnesses. Furthermore, CL leads to viewing S3 and S4 as evidence against puneria and for zymosis, whereas NCL results in viewing S3 as evidence for both illnesses, and S4 as nondiagnostic of puneria but evidence for zymosis. In short, CL blurs the distinction between evidence for one illness and evidence against the other, whereas NCL encourages the distinction. The result is that CL leads to dependent confidence in the illnesses and NCL leads to independent confidence.

After completing the learning phase, subjects were told that they had become specialists who see patients known to have a class of illness, of which there were only two types: puneria and zymosis. All subsequent patients were known to have one illness or the other, but not both. Subjects could not proceed to the target tasks until they understood that the illnesses were mutually exclusive and exhaustive, and subjects were reminded of this before each target task. They were also asked about their understanding of the relation between the illnesses after the target tasks were completed. When presented with four options (mutually exclusive only, exhaustive only, both, neither), virtually every subject chose the correct answer. Furthermore, subjects were told that they would be reporting confidence that a patient had one illness or the other, and their judgments should reflect expected percent correct. That is, on occasions when they reported, for example, 90% confidence, they should expect the patient to have the focal illness 90% of the time. Subjects responded on a scale of 1 to 99.

The first target task was a belief-updating task. Subjects were presented with individual symptoms and updated confidence in one illness for various patients. They then repeated the task with the other illness focal. No feedback was provided. It was found that CL subjects' confidence tended to change in a complementary manner, whereas NCL subjects' changes tended to be noncomplementary (McKenzie, 1997, Experiment 1). For example, when S1 was presented, CL subjects' confidence increased when puneria was focal, and decreased when zymosis was focal. In contrast, NCL subjects increased confidence in puneria, but did not change confidence in zymosis. Similarly, CL subjects increased confidence in puneria and decreased confidence in zymosis when presented with S2. NCL subjects, however, increased confidence in both illnesses. Analogous differences between CL and NCL subjects occurred for S3 and S4 as well. All these differences are as predicted by the dependent/independent confidence distinction and show that NCL subjects' confidence was influenced by evidence for and against the focal illness, but not by evidence for and against the alternative.

The target task of more direct interest was the second one, in which subjects were presented with complete patient profiles and reported confidence in one of the illnesses (there was again no feedback). All 16 (2^4) unique profiles were presented with one illness focal, then again with the other illness focal. Table 6 shows the profiles and the normative (Bayesian) probability of each illness,

TABLE 6

The 16 Profiles, Accompanying Symptoms, and Probability of Each Illness Given the Symptoms

Profile	S1 S2 S3 S4	p(puneria)	p(zymosis)
1	Pr Pr Pr Pr	.50	.50
2	Pr Pr Pr Ab	.94	.06
3	Pr Pr Ab Pr	.69	.31
4	Pr Pr Ab Ab	.97	.03
5	Pr Ab Pr Pr	.31	.69
6	Pr Ab Pr Ab	.88	.12
7	Pr Ab Ab Pr	.50	.50
8	Pr Ab Ab Ab	.94	.06
9	Ab Pr Pr Pr	.06	.94
10	Ab Pr Pr Ab	.50	.50
11	Ab Pr Ab Pr	.12	.88
12	Ab Pr Ab Ab	.69	.31
13	Ab Ab Pr Pr	.03	.97
14	Ab Ab Pr Ab	.31	.69
15	Ab Ab Ab Pr	.06	.94
16	Ab Ab Ab Ab	.50	.50

Note. Pr and Ab refer to "present" and "absent," respectively.

given the symptoms. CL subjects' confidence in the two illnesses was largely additive, while NCL subjects' confidence was not (McKenzie, in press, Experiment 3). For example, when presented with Profile 1, which has all four symptoms present and is consistent with both illnesses, CL subjects reported about 50% confidence in each illness; thus, these judgments summed to about 100%, indicating additivity. However, NCL subjects reported about 80% confidence in each illness for that profile. For Profile 16, which has no symptoms present and is consistent with neither illness, CL subjects again reported about 50% confidence in each illness, but NCL subjects were only about 15% confident in each. These differences in additivity are as predicted by the different representations of confidence. CL subjects' confidence in the focal illness was a function of how well the symptoms were consistent with each illness, whereas NCL subjects' confidence was based on how consistent the symptoms were with the focal illness without regard to the alternative. In other words, CL subjects took into account the strength of the alternative, whereas NCL subjects did not.

On the surface, a potential alternative explanation is that NCL subjects (despite denying this when asked directly) responded as though patients could have neither illness in addition to puneria or zymosis. Indeed, this could explain why their judgments summed to only about 30% for Profile 16: 70% confidence could have been allocated to the third "neither illness" hypothesis that was not asked about. However, this cannot explain these same subjects' superadditive judgments for Profile 1: Judgments summed to about 160%, so adding a third hypothesis cannot help confidence sum to 100%.

Another potential concern is that NCL subjects saw 120 healthy patients in addition to the 120 puneria and zymosis patients. Perhaps NCL subjects attended less well to puneria and zymosis patients because they comprised only half the trials. However, NCL subjects' judgments were neither random nor simply noisier than CL subjects' judgments. The pattern of responses for each group was clear and as predicted.

It is worth noting that the massive and systematic nonadditivity on the part of NCL subjects is unprecedented, as far as I know. Nonadditivity has been reported before, but virtually always in the form of superadditivity for three or more hypotheses (Robinson & Hastie, 1985; Teigen, 1983; Van Wallendael, 1989; Van Wallendael & Hastie, 1990; Wright & Whalley, 1983). Here, however, sub- and superadditivity was found for two hypotheses. These findings call into question any descriptive theory of subjective probability that assumes or implies additivity for two mutually exclusive and exhaustive hypotheses (e.g., Tversky & Koehler, 1994).

REANALYSIS OF MCKENZIE (IN PRESS), EXPERIMENT 3

The question now is whether more overconfidence results from an independent representation, where the strength of the alternative tends to be underweighted or ignored, than from a dependent representation, where the alternative tends to be taken into account. To answer this, results were reanalyzed for the target task in which complete patient profiles were presented and subjects reported confidence in each illness (on separate occasions).

Results and Discussion

The 38 CL and 37 NCL subjects each reported 32 probabilities (16 patients $\times$ 2 illnesses), which were divided by 100 for these analyses, putting them on a scale of .01 to .99. Table 1 (middle rows) shows performance on the same measures discussed for the simulations. Though (half-scale) mean confidence was similar for both groups ($t(73) = 1.05$, $p = .30$), E(correct) was lower for NCL subjects ($t = 3.77$, $p < .001$), resulting in greater overconfidence ($t = 2.38$, $p = .02$). Furthermore, *MSD* was higher for NCL subjects, indicating poorer judgment accuracy ($t = 3.5$, $p < .001$). Finally, the higher *MAD* shows that NCL subjects' confidence judgments were less additive than CL subjects' judgments ($t = 8.39$, $p < .0001$). (Additivity results are reported in McKenzie, in press, but are presented here for completeness.)

These results are generally consistent with the pattern generated by the computer simulations. NCL subjects underweight the strength of the alternative hypothesis, leading to decreased performance on several measures, including increased overconfidence. The only exception was that NCL subjects' confidence was not higher than CL subjects' confidence; in fact, it was slightly lower. One reason for this might be that NCL subjects knew that they were in a somewhat difficult situation, having to apply what they had learned to a new setting, and (appropriately) reduced their confidence. Nonetheless, the reduction was insufficient, leading to more overconfidence relative to CL subjects.

REANALYSIS OF MCKENZIE (IN PRESS), EXPERIMENT 4

Under what conditions might NCL subjects be more inclined to take into account the strength of the alternative? In the above experiment, subjects were presented with patient profiles and asked asymmetric questions, such as "How confident are you that this patient has puneria?" The alternative illness was not mentioned. Asking symmetric questions, such as "How confident are you that this patient has puneria rather than zymosis?" should encourage subjects to take into account the alternative (see, e.g., Baron *et al.*, 1988; Beyth-Marom & Fischhoff, 1983; Zuckerman *et al.*, 1995; see also Tweney *et al.*, 1980). Because CL subjects already take into account the alternative, question type should have no effect on their behavior. NCL subjects, however, tend to underweight the alternative, so symmetric questions should lead them to consider the alternative more. This is the case (McKenzie, 1997, Experiment 2, in press, Experiment 4). In the task presenting subjects with complete profiles, NCL subjects were almost 90% confident in each illness when presented with Profile 1 and asked asymmetric questions, but were less than 70% confident in each with symmetric questions. Similarly, when presented with Profile 16, NCL subjects were about 15% confident in each illness with asymmetric questions, but were about 30% confident in each with symmetric questions. (Note that these shifts in confidence are consistent with Eq. (4) assuming that symmetric questions increased *w*.) CL subjects were about 50% confident in each illness for both profiles regardless of question type. Symmetric questions did not eliminate differences between CL and NCL subjects, but they were considerably reduced. The important point is that symmetric questions led to more consideration of the alternative for NCL but not CL subjects.

Regarding the current dependent measures of interest, the simulation results lead to the following predictions: CL subjects should not be affected by question type, but for NCL subjects, symmetric questions should lead to decreased confidence, increased *E*(correct), and decreased overconfidence, *MSD,* and *MAD*. These predictions were tested through reanalyzing the results of McKenzie, in press, Experiment 4. Half the CL and NCL subjects in this experiment were asked asymmetric questions (*N*s = 44), and half were asked symmetric questions (*N*s = 42).

Results and Discussion

The ANOVA (learning × question type) results for each of the five measures are shown in Table 7. The main reason for presenting them is to show that the effects of learning were again present. The only difference is that there is now a main effect of learning on confidence, with CL subjects exhibiting higher confidence (see the discussion of the first reanalysis). Table 7 also shows that there were main effects of question type on all dependent measures except *E*(correct), and that there was an interaction between learning and question type only for the measure of nonadditivity, *MAD*.

The bottom portion of Table 1 shows the cell means for each measure. Despite the general lack of interactions in the ANOVAs, contrasts examining the effect of question type on each learning group separately revealed a clear trend. For CL subjects, question type had no effect on any dependent measure (all ps > .16). For NCL subjects, however, symmetric questions led to lower confidence ($t(84) = 2.24$, $p = .028$), overconfidence ($t = 3.23$, $p = .002$), *MSD* ($t = 2.43$, $p = .02$), and *MAD* ($t = 5.2$, $p < .0001$). The only measure not significantly affected by question type for NCL subjects was *E*(correct) ($p = .13$), though the effect was in the predicted direction.

TABLE 7

Analyses of Variance for Reanalysis of McKenzie (in press), Experiment 4

	F				
Source	Confidence	*E*(Correct)	Overconfidence	*MSD*	*MAD*
Learning (L)	21.66**	49.96**	4.32*	31.63**	86.94**
Q Type (Q)	4.68*	3.39	11.14**	5.80*	22.17**
L × Q	1.23	0.38	2.08	1.25	13.15**
Error	(0.007)	(0.007)	(0.010)	(0.003)	(0.015)

*$p < .05$.
**$p < .01$.
Note. Values in parentheses represent mean square errors. $df = 1{,}168$.

The pattern of results was largely as predicted from the computer simulations: Because CL subjects take into account the alternative, question type had no effect on any of the measures. However, symmetric questions improved NCL subjects' performance on virtually every measure, most notably overconfidence. The fact that the same pattern of differences was found between the two NCL groups is important because, unlike the CL versus NCL comparison, question type is the only difference between these groups. Asking a question that encouraged consideration of the alternative led to the predicted changes on virtually all the dependent measures, even though the learning history, hypotheses, and evidence were identical for the two NCL groups.

GENERAL DISCUSSION

The notion that subjects underweight evidence regarding the alternative has been much discussed in the overconfidence literature, but, as pointed out earlier, this can mean at least three things. The emphasis here was on consideration of evidence for and against the focal hypothesis while underweighting or ignoring evidence for and against the alternative. Computer simulations showed that, as the strength of the alternative was underweighted, confidence increased and proportion correct decreased, leading to more overconfidence; furthermore, judgment accuracy and additivity decreased. Additional simulations examining the robustness of these findings indicated that underweighting the alternative leads to the same pattern of results under quite general conditions. Reanalyses of data from two experiments, in which one group of subjects was known to take into account the strength of the alternative and one was known to underweight it, revealed a similar pattern. Furthermore, encouragement to take into account the alternative led to changes in overconfidence and other behavior for only the latter group, and these changes were largely as predicted by the simulations.

These findings seem to be the most direct yet regarding this issue. Most studies of overconfidence use hypotheses for which there is no known objective probability. More important is that it is usually virtually impossible to know what evidence subjects are using, let alone how they are using it. In the empirical studies discussed here, however, all these factors were known.

The claim is not that underweighting the strength of alternatives is the sole cause of overconfidence, which is almost certainly an overdetermined phenomenon. Indeed, even subjects in the reanalyzed experiments who took into account the alternative exhibited considerable overconfidence. However, the current results—the simulation results, in particular—showed that underweighting the alternative is sufficient for overconfidence (given reasonable assumptions). Underweighting alternatives is ubiquitous in laboratory tasks that can even loosely be construed as hypothesis testing (e.g., Evans, 1989; Fischhoff & Beyth-Marom, 1983; Klayman & Ha, 1987; McKenzie, 1994). To the extent that a given task investigating overconfidence is one of testing hypotheses, one might expect the alternative(s) to be underweighted, and that this would be at least partially responsible for overconfidence.

However, the distinction between dependent and independent confidence effects two important caveats to the above statement. First, dependent confidence leads to consideration of the alternative. How confidence in hypotheses is represented is an important issue for the study of (over)confidence. The key notion appears to be whether there is a distinction between evidence for one hypothesis and evidence against the other. If there is no distinction, then confidence is dependent and the strength of the alternative will be taken into account. For example, if one were predicting rain, the alternative is that it will not rain, and there is no distinction between evidence for the "rain" hypothesis (e.g., dark clouds) and evidence against the "no rain" hypothesis; they are the same thing. The same evidence is deemed relevant regardless which hypothesis is focal. In contrast, a distinction can be drawn between evidence for absinthe being a liqueur and evidence against it being a precious stone. An example of the former is recalling absinthe being offered in a restaurant or bar; an example of the latter is being somewhat familiar with precious stones, but not recalling absinthe being one of them. The first piece of evidence might be weighted more heavily when the "liqueur" hypothesis is focal, the second when the "precious stone" hypothesis is focal. The (psychological) existence of the "evidence for one/against the other" distinction implies independent confidence in the two hypotheses, a greater likelihood that the alternative will be underweighted, and increased overconfidence.

The second caveat is that task variables that encourage taking into account the alternative lead subjects with independent confidence to behave more similarly to subjects with dependent confidence. It was mentioned earlier that McKenzie (1997, Experiment 1) found that subjects with independent confidence updated beliefs in a noncomplementary fashion. Symmetric questions decreased noncomplementarity, but did not eliminate it (Experiment 2). However, asking contiguous questions (i.e., asking about both competing hypotheses one after the other rather than on separate occasions) that were symmetric led to complementary updating for these subjects (Experiment 3). Qualitative differences between the groups were eliminated.

In sum, then, the present view can be seen as one concerned with how cognition and task features interact to influence overconfidence. The claim is not that subjects generally underweight alternatives; instead, an important factor is how confidence in the competing hypotheses is cognitively represented. Furthermore, though underweighting the alternative is more likely with independent rather than dependent confidence, differences in performance can be reduced, possibly even eliminated, depending on task variables such as how confidence is probed.

REFERENCES

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

Arkes, H. R., & Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General,* **112,** 117–135.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes,* **42,** 88–110.

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology,* **45,** 1185–1195.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance,* **7,** 928–935.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica,* **43,** 111–121.

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

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

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

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

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

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

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.

Klayman, J., & Brown, K. (1993). Debias the environment instead of the judge: An alternative approach to reducing error in diagnostic (and other) judgment. *Cognition,* **49,** 97–122.

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

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **20,** 461–469.

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

Lipe, M. G. (1990). A lens model analysis of covariation research. *Journal of Behavioral Decision Making,* **3,** 47–59.

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. (1997). *(Non)Complementarity when updating beliefs in two hypotheses.* Manuscript submitted for publication.

McKenzie, C. R. M. (in press). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

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.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes,* **40,** 193–218.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General,* **110,** 101–120.

Sniezek, J. A., Paese, P. W., & Switzer, F. S., III (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes,* **46,** 264–282.

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.

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.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., & Arkkelin, D. L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology,* **32,** 109–123.

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 and Cognition,* **18,** 240–250.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology,* **19,** 231–241.

Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **16,** 509–521.

Wright, G., & Whalley, P (1983). The supra-additivity of subjective probability. In B. P. Stigum & F. Wenstop (Eds.), *Foundations of utility and risk theory with applications* (pp. 233–244). Reidel: Dordrecht, Holland.

Zuckerman, M., Knee, C. R., Hodgins, H. S., & Miyake, K. (1995). Hypothesis confirmation: The joint effect of positive test strategy and acquiescence response set. *Journal of Personality and Social Psychology,* **68,** 52–60.

Received: March 26, 1997