Organizational Behavior and Human Decision Processes
Vol. 81, No. 2, March, pp. 195–225, 2000
doi:10.1006/obhd.1999.2876, available online at http://www.idealibrary.com on IDEAL®

The Binary Additivity of Subjective Probability Does not Indicate the Binary Complementarity of Perceived Certainty

Paul D. Windschitl

University of Iowa

People's numeric probability estimates for 2 mutually exclusive and exhaustive events commonly sum to 1.0, which seems to indicate the full complementarity of subjective certainty in the 2 events (i.e., increases in certainty for one event are accompanied by decreases in certainty for the other). In this article, however, a distinction is made between the additivity of probability estimates and the complementarity of internal perceptions of certainty. In Experiment 1, responses on a verbal measure of certainty provide evidence of binary noncomplementarity in the perceived likelihoods of possible scenario outcomes, and a comparison of verbal and numeric certainty estimates suggests that numeric probabilities overestimated the complementarity of people's certainty. Experiment 2 used a choice task to detect binary noncomplementarity. Soliciting numeric probability estimates prior to the choice task changed the participants' choices in a direction consistent with complementarity. Possible mechanisms yielding (non)complementarity are discussed. © 2000 Academic Press

When people are asked about the probabilities of three or more mutually exclusive and exhaustive events, their estimates for the events often greatly exceed 1.0 (e.g., Robinson & Hastie, 1985; Teigen, 1974, 1983; Tversky & Koehler, 1994; Van Wallendael, 1989; Van Wallendael & Hastie, 1990; Wright & Whalley, 1983). This nonadditivity of subjective probabilities violates a normative standard of probability. However, when people are asked about two mutually exclusive and exhaustive events, their subjective probabilities typically sum to about 1.0. This binary additivity has been observed in numerous studies across a variety of event domains (e.g., Wallsten, Budescu, & Zwick, 1993;

I thank Irwin Levin for his helpful comments on a draft of this manuscript.

Address correspondence and reprint requests to Paul D. Windschitl, Department of Psychology, University of Iowa, Iowa City, Iowa, 52242. E-mail: paul-windschitl@uiowa.edu.

Teigen, 1983; Tversky & Fox, 1994; Tversky & Koehler, 1994). Not only is binary additivity in line with normative models of probability, it is also a key aspect of a major descriptive model of how people make judgments of probability (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994).

What do demonstrations of binary additivity suggest about people's perceptions of certainty? It is commonly assumed that the binary additivity of subjective probabilities indicates the near perfect complementarity of people's certainty in binary cases. That is, as certainty in one event increases, certainty in the other will decrease appropriately. In this article, an alternative interpretation is offered—one that does not assume that the additivity of people's probability estimates reflects the complementarity of perceived certainty. This argument is based on a distinction between subjective probabilities and the underlying construct of certainty.

SUBJECTIVE PROBABILITY AND PSYCHOLOGICAL CERTAINTY

A common assumption in research on judgment and decision making is that numeric measures of subjective probability provide an accurate method of assessing people's perceptions of certainty. Although it is true that measures of subjective probability are often effective at assessing psychological certainty, it is important to recognize that subjective probabilities are not direct representations of people's perceptions of certainty and that they are in some respects highly artificial representations (see Windschitl & Wells, 1996). For most everyday judgments and decisions under uncertainty, people do not perform numeric calculations and compare subjective probabilities to numeric thresholds. Formal numeric probability systems were not developed until the 17th century (Zimmer, 1983), and it seems unlikely this modern innovation has replaced whatever systems humans used to handle uncertainty up until that point. It seems reasonable to assume that much of how people typically think about uncertainty is pre-Bernoullian.

Windschitl and Wells (1996) proposed that soliciting numeric probability estimates from people prompts them to make considerations that they would not normally make when forming impressions of certainty in many judgment and decision-making situations. For example, soliciting numeric probability estimates from subjects may enhance their concerns about the accuracy of their responses. Subjects are aware that the accuracy of a numeric estimate—unlike most decisions and nonnumeric judgments—can be readily compared to a normative standard. Also, soliciting numeric estimates might prime subjects' awareness of the applicability of formal rules. Whether or not subjects know how to use the primed rules, they might believe that the rules are relevant and attempt to generate responses that are consistent with their understanding of the rules.

In support of these claims, several studies have shown that alternatives to the traditional subjective probability measures can be sensitive to a variety of manipulations affecting psychological certainty, even though subjective probability measures are insensitive to the manipulations (Kirkpatrick & Epstein, 1992; Windschitl & Martin, 1999; Windschitl & Weber, 1999; Windschitl & Wells, 1996, 1998). For example, in demonstrations of the alternative-outcomes effect, Windschitl and Wells (1998) used nonnumeric measures of uncertainty as well as a decision measure to show that certainty in a focal outcome (e.g., you winning a raffle in which you hold 22 tickets) is partially a function of how alternative outcomes are distributed (many of the 56 other tickets are held by one person versus distributed evenly among several people). Although judgments and decisions were sensitive to the alternative-outcomes manipulations, subjective probabilities were insensitive to the manipulations and tended to conform to normative standards. This indicates that participants were more likely to utilize their understanding of formal rules of probability when generating numeric probability estimates than when making other types of judgments or decisions mediated by uncertainty.

SUBJECTIVE PROBABILITY AND BINARY COMPLEMENTARITY

The present work is based on a related observation about measures of subjective probability. Generating a subjective probability estimate causes people to scale their own levels of certainty in terms of probabilities. This would prompt an awareness of the constraints of the probability scale. Specifically, the scale itself prompts an awareness of the additivity and complementarity constraints. Once these constraints are recognized by a respondent, it is relatively easy to conform to additivity and complementarity in binary cases (but less so in nonbinary cases). For example, most research participants know that assigning a chance estimate of 30% to a focal hypothesis leaves the remaining 70% for the alternative hypothesis, and if 70% is too great of a chance for the alternative, then the estimate for the focal hypothesis must be changed. Participants are essentially faced with a task of partitioning a 100-point scale (or 0 to 1 scale) among the two hypotheses. This task draws attention to the alternative hypothesis and boosts the likelihood that its support (i.e., the amount of evidence supporting the hypothesis) is assessed along with the support for the focal hypothesis. Hence, soliciting numeric probability estimates will increase the chance that evidence relevant to each of the two hypotheses will receive equal weight in the judgment process. If this occurs, binary complementarity will be observed.

That subjective probability estimates will conform to binary complementarity is a main tenet of support theory, a recently proposed theory of subjective probability (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994). According to support theory, the subjective probability of a focal hypothesis (A) rather than its alternative (B) can be represented as:

$$P(\mathrm{A,B}) = \frac{\mathrm{s(A)}}{\mathrm{s(A)} + \mathrm{s(B)}}$$

One consequence of this equation is that as *support* for the alternative hypothesis increases, the subjective probability of the focal hypothesis will decrease accordingly (assuming constant support for the focal hypothesis). This describes perfect complementarity.

Although numeric probability estimates may tend to exhibit binary complementarity, I argue that internal assessments of certainty are less likely to exhibit this property. For most judgments, decisions, and behaviors under uncertainty, people do not spontaneously scale their own certainty in terms of numeric subjective probability. That is, they generally aren't thinking numerically when they think about the likelihood that their illness was due to food poisoning not a virus, that Candidate X would be hired rather than Candidate Y, that they could get a better parking space in the next lot, or that a computer problem is due to software rather than hardware. Hence, concerns with additivity and complementarity are not necessarily activated in everyday judgment tasks, and the support for the alternative hypothesis might not receive as much attention as it would if numeric probabilities were solicited. This would cause the evidence that is directly relevant to the focal hypothesis to receive more weight than evidence relevant to the alternative hypothesis, and perfect complementarity would not follow.

In regard to support theory, this argument suggests that although the expression s(A)/[s(A) + s(B)] describes the numeric subjective probability estimate for a focal hypothesis (just as the theory purports), it does not necessarily describe the internal assessment of certainty that drives behavior. When judging certainty for purposes of engaging in a behavior, people do not spontaneously derive a probability estimate by normalizing s(A) across s(A) + s(B). Rather, s(A) can have a more direct influence on certainty and behavior, while s(B) is underweighted or ignored.

In sum, the preceding arguments yield two assertions. First, full binary complementarity is not a general property of people's certainty. Second, measures of subjective probability overestimate the degree to which people's internal perceptions of certainty conform to complementarity.

Recent work by McKenzie (1998) has provided some initial yet compelling evidence for the first assertion (see also McKenzie, 1999). Participants in that research learned about symptoms for two fictitious diseases: zimosis and puneria. *Contrastive learners* underwent a training phase that essentially taught them whether each of a series of symptoms was diagnostic of zimosis or puneria. *Noncontrastive learners* underwent a training phase that taught them whether each of a series of symptoms was or was not diagnostic of zimosis and was or was not diagnostic of puneria. After the learning phase, participants were shown sets of symptoms for hypothetical patients and were told that the patients suffered from one of the two illnesses but not both. Each participant provided a probability estimate for one illness and later an estimate for the other illness. While the estimates for the two illness were largely additive for the contrastive learners, the estimates showed predicted patterns of subadditivity and superadditivity for noncontrastive learners. When the patient had symptoms that fit both diseases, combined estimates for the two diseases tended to exceed 1.0. When the patient had symptoms that fit neither disease, estimates for the two diseases tended to fall below 1.0.

McKenzie's (1998) experiments not only provide an important and compelling demonstration that psychological certainty can violate binary complementarity, but they also established this point with subjective probabilities. To achieve this demonstration, a highly structured learning paradigm was utilized. The present experiments took a different approach to testing for violations of binary complementarity. It was assumed that observing violations of binary complementarity does not require a highly controlled learning paradigm. Rather, it was assumed that violations of binary complementarity might be readily observed within a scenario paradigm if nontraditional measures of psychological certainty were used. Such observations would add unique and broader evidence for the first assertion that full binary complementarity is not a general property of people's perceptions of certainty.

The present experiments also tested the second assertion that subjective probabilities overestimate the degree to which people's perceptions of certainty exhibit complementarity. This assertion calls into question the assumption that the additivity of subjective probabilities is a good standard for drawing conclusions about the complementarity of perceived certainty.

EXPERIMENT 1

If binary complementarity is not a general property of psychological certainty, then it should be possible to increase people's certainty in one of two mutually exclusive and exhaustive (MEE) hypotheses without reducing their certainty in the other. It was assumed that the key to testing for such an effect is measuring psychological certainty without asking people to scale their certainty on a numeric probability scale. Hence, Experiment 1 utilized a verbal certainty scale. This scale, shown in Appendix A, requires that participants map their perceptions of certainty onto 1 of 11 verbal certainty phases. For comparison purposes, Experiment 1 also utilized a numeric subjective probability scale. This scale, also shown in Appendix A, required that participants map their perceptions of certainty onto 1 of 11 numeric estimates.

In most studies addressing complementarity, complementarity is actually assessed through additivity; when subjective probability estimates add to 1.0, full complementarity is assumed. The verbal scale used in this research, however, does not allow for such an analysis. There is no reason to assume that the responses made on the verbal scale should or can be meaningfully translated into numeric probabilities (for discussion, see Windschitl & Wells, 1996), so it is unclear when a set of verbal certainty estimates should be considered additive versus nonadditive. Hence, complementarity was assessed without reference to additivity.

Participants read four scenarios that defined a set of MEE hypotheses and provided information about them. Two of the scenarios described sets of two MEE hypotheses—thus allowing for tests of binary complementarity. To allow for tests of complementarity in nonbinary cases, the other two scenarios described sets of three and four MEE hypotheses. Participants read either a strong-evidence version or a weak-evidence version of each scenario. For weak-evidence versions, none of the possible hypotheses received much support from the described evidence, whereas for the strong-evidence versions, all of the possible hypotheses received support from the evidence.

For example, the Politician-of-the-Year Scenario indicated that there were two candidates remaining for a political award. In the weak-evidence version, neither of the two candidates fit the presumed stereotype of a candidate who would win the type of award that was described. The award was for “Republican of the Year.” One candidate, Rebecca Sharp, was instrumental in running campaigns but was disliked by the “most powerful male and older Republicans,” and the other candidate, Stacy Rakan, raised large amounts of money but was not favored by conservatives because of her attempts to keep abortion legal. In the strong-evidence version, both candidates appeared as strong contenders for a “Democrat of the Year Award.” Rebecca Sharp was recently elected to various offices and encouraged to run for the Senate. Stacy Rakan raised large amounts of money and was encouraged to run for a House of Representatives Seat.

It was expected that for participants reading the weak-evidence version, initial perceptions of certainty in both candidates would be low, but for participants reading the strong-evidence version, certainty in both candidates would be high. This pattern would clearly violate the constraints of complementarity. Note, however, that this expected pattern concerns psychological certainty, not subjective probability. For those participants asked to provide numeric subjective probability estimates, their perceptions of certainty would change. The solicitation of a numeric probability would force participants to partition their certainty on a 100-point scale, and they would be prompted to compare the evidence for the two candidates (more so than they would have had they not been asked for a numeric response). The result would be full complementarity. For example, a participant who decides that Rebecca Sharp has a 90% chance of winning would probably realize that assigning a 60% chance to Stacy Rakan would violate the constraints of the scale.

For those participants asked to provide verbal certainty estimates, a similar process might occur, but to a significantly lesser extent. Although the solicitation might cause some participants to view their task as a partitioning of certainty, the softer constraints of the verbal scale would still allow noncomplementarity to emerge. For example, a participant who decides that Rebecca Sharp is extremely likely to win could also indicate that Stacy Rakan is fairly likely to win without violating some hard constraint of the scale.

Finding that a participant gave responses of “extremely likely” and “fairly likely” to the two candidates suggests noncomplementarity, but it is somewhat problematic evidence because it is impossible to precisely determine when a set of verbal responses is additive or not additive.

Hence, a key prediction for Experiment 1 was that overall levels of certainty expressed on the verbal scale—derived by summing the responses for the set of hypotheses in a scenario—would be significantly greater in the strong-evidence versions than in the weak-evidence versions. This would indicate that increases in certainty for one hypothesis were not accompanied by fully complementary changes in certainty for the other hypothesis(es). Although some noncomplementarity was also anticipated for numeric responses in nonbinary cases, the degree of noncomplementarity was expected to be significantly greater for verbal versus numeric responses in all scenarios. Also, numeric responses were expected to exhibit nearly perfect complementarity and additivity for the two scenarios involving binary cases.

Method

Participants and design. The participants were 440 undergraduate students at Iowa State University who received credit in an introductory psychology course. The design was a 2 (evidence strength) × 2 (scale type) between-subjects factorial.

Scenarios. Two versions (strong- and weak-evidence version) of four scenarios were constructed for the experiment. These scenarios are shown in Appendix B. Each scenario described a set of MEE hypotheses (2, 3, or 4 hypotheses) and provided evidence relevant to each hypothesis. In the strong-evidence version, there was evidence that supported all hypotheses. In the weak-evidence version, the evidence was not very supportive of any of the hypotheses. At the end of each scenario, there were questions that asked participants about their certainty for each hypothesis described in the scenario. The questions were accompanied by either numeric or verbal response scales.

Scales. The numeric and verbal scales used in Experiment 1 (shown in Appendix A) are based on those introduced by Windschitl and Wells (1996). Their method of ordering of the responses options on the verbal scale was based on pilot testing in which participants translated verbal expressions into numeric probabilities. For example, the median responses for “certain,” “extremely likely,” and “quite likely” were 99, 90, and 80%, respectively. Although those translation data seem to suggest that the verbal response options should approximate the numeric response options, this type of equivalence should not be assumed (see Windschitl & Wells, 1996). Therefore, responses on the verbal scale are not compared directly to responses on the numeric scale. That is, although the effects of the evidence-strength manipulation on the verbal scale are compared to those effects on the numeric scale, overall scores or individual means on the verbal scale are not compared to those on the numeric scales.

Results

Responses to each of the certainty questions were scored from 0 to 10 for both the numeric and verbal scales (0 = “0%” or “impossible”; 10 = “100%” or “certain”). Several alternative methods for scoring verbal responses were also

TABLE 1

Overall-Certainty Scores for Each Version of the Scenarios in Experiment 1

Scenario/version	Numeric scale *M* (*SD*)	Verbal scale *M* (*SD*)
Politician-of-the-Year		
Weak-evidence	9.9 (1.5)	10.3 (2.2)
Strong-evidence	10.5 (1.6)	12.4 (2.4)
Profession		
Weak-evidence	10.0 (1.1)	11.3 (2.0)
Strong-evidence	10.5 (1.7)	12.0 (2.3)
Senate Race		
Weak-evidence	11.5 (2.9)	16.9 (3.3)
Strong-evidence	12.6 (4.1)	19.2 (3.0)
Four Suspects		
Weak-evidence	13.2 (4.5)	19.4 (4.2)
Strong-evidence	13.2 (5.2)	21.5 (3.9)

Note. The overall-certainty scores were generated by summing a participant's responses for all hypotheses in a given scenario.

tested, all of which yielded essentially the same results.[1] The analyses reported below are based on *overall-certainty scores*, which were computed by summing across a given participant's responses to the MEE hypotheses described in a scenario. For example, if a participant's responses for the two politicians in the Politician-of-the-Year Scenario were 40 and 70%, then his/her overall-certainty score was 11 (4 + 7). Table 1 presents the means for the overall-certainty scores from each scenario. The means for the individual hypotheses of each scenario can be found in Appendix C.

Before describing the results at the scenario level, I summarize the overall findings. The two assertions described above received strong support. First, for all of the scenarios, including the two scenarios describing binary cases, participants providing verbal responses exhibited noncomplementarity; the overall-certainty scores from those scenarios were significantly higher for the

[1] The question of how to quantify the responses on a verbal scale is a complex one (Windschitl & Wells, 1996). Instead of focusing on this question here, I tested several of the most plausible methods for scoring the verbal responses and found that all of these methods led to the same conclusions regarding the issues at hand. The described scoring method, which treats the verbal responses as equidistant (*impossible* = 0, *extremely unlikely* = 1, *quite unlikely* = 2, etc.), yields essentially the same results as the following methods: a method that assumes that responses in the middle of the scale are less distinct than responses near the ends (0, 1.25, 2.5, 3.75, 4.5, 5, 5.5., 6.25, 7.5, 8.75, 10), a method that assumes that responses at the ends of the scale are less distinct than responses in the middle (0, 0.5, 1.25, 2.25, 3.5, 5, 6.5, 7.75, 8.75, 9.5, 10), a method that assumes that responses at the low end of the scale are less distinct than responses at the high end (0, 0.5, 1, 1.75, 2.5, 3.5, 4.5, 5.75, 7, 8.5, 10), a method that assumes that responses at the high end of the scale are less distinct than responses at the low end (0, 1.5, 3, 4.25, 5.5, 6.5, 7.5, 8.25, 9, 9.5, 10), and a method that assumes that responses near *impossible, as likely as unlikely*, and *certain* would be more distinct than responses that are not near these natural anchor points (0. 1.25, 2.25, 2.75, 3.75, 5, 6.25, 7.25, 7.75, 8.75, 10). Hence, the manner in which the verbal responses were quantified does not account for the key findings of Experiment 1.

strong-evidence versions than for the weak-evidence versions. Second, for three of the four scenarios, the evidence-strength manipulation had a stronger impact on overall-certainty scores from verbal responses than from numeric responses. In other words, the verbal measure detected a significantly greater degree of noncomplementarity than did the numeric measure. If one assumes that effects detected with the verbal measure reflect underlying differences in internal perceptions of certainty, then one can conclude that the numeric measure underestimated the noncomplementarity of perceived certainty.

Politician-of-the-Year Scenario. As predicted, the overall-certainty scores for participants giving verbal responses were significantly higher in the strong-evidence condition than in the weak-evidence condition, $t(219) = 7.06$, $p < .001$. This large effect ($d = .91$)[2] indicates binary noncomplementarity; if high certainty in one candidate had always been appropriately balanced with low certainty in the other candidate (and vice versa), then overall-certainty scores would have been equivalent in the strong- and weak-evidence conditions. Regarding the numeric responses, inspection of overall-certainty means in Table 1 suggests that those responses were largely additive and exhibited complementarity. However, there was a small but significant trend for overall-certainty scores to be higher in the strong-evidence version than in the weak-evidence version, $t(217) = 3.16$, $p < .01$, $d = .39$. This suggests that, even with numeric subjective probabilities, some degree of binary noncomplementarity can be observed with this paradigm. Finally, the prediction that the evidence-strength manipulation would have a stronger effect on overall certainty expressed on the verbal scale than on the numeric scale was supported by a significant Evidence-Strength × Response-Scale interaction, $F(1, 436) = 16.00$, $p < .001$. This finding suggests that noncomplementarity was underestimated by the numeric scale.

Profession Scenario. As predicted, the overall-certainty scores for verbal responses were significantly higher in the strong-evidence condition than in the weak-evidence condition, $t(219) = 2.38$, $p < .05$, $d = .32$. This effect is smaller than the one detected in the Politician-of-the-Year Scenario, but nonetheless it indicates binary noncomplementarity. Regarding the numeric responses, there was again a small but significant trend for overall-certainty scores to be higher in the strong-evidence version, $t(217) = 2.42$, $p < .05$, $d = .35$. Finally, unlike the Politician-of-the-Year Scenario, there was no evidence that the strength manipulation had a stronger effect on overall certainty expressed on the verbal scale than on the numeric scale; the Evidence Strength × Response Scale interaction was not significant, $F(1, 436) = 1.55$, $p = .50$.

Senate-Race Scenario. For this scenario involving three MEE hypotheses, the overall-certainty scores for verbal responses were again significantly higher in the strong-evidence condition than in the weak-evidence condition,

[2] This d statistic refers to the difference between means in standard deviation units. Cohen (1988) has suggested that effects with magnitudes of .20, .50, and .80 should be considered “small,” “medium,” and “large” effects, respectively.

$t(219) = 5.34$, $p < .001$, $d = .73$. Regarding the numeric responses, there was a significant trend for overall-certainty scores to be higher in the strong-evidence version, $t(217) = 2.22$, $p < .05$, $d = .31$; such a finding is not unexpected for cases with more than two MEE hypotheses. Finally, the Evidence Strength × Response Scale interaction was marginally significant, again suggesting that noncomplementarity was underestimated by the numeric scale, $F(1, 436) = 3.49$, $p = .06$.

Four-Suspects Scenario. The overall-certainty scores for verbal responses were again significantly higher in the strong-evidence condition than the weak-evidence condition, $t(219) = 3.92$, $p < .001$, $d = .52$. Somewhat surprisingly, there was no difference in overall numeric scores between those conditions, $t < 1$. Finally, a significant Evidence Strength × Response Scale interaction again suggests that noncomplementarity was underestimated by the numeric scale, $F(1, 436) = 6.33$, $p < .05$.

Discussion

The fact that noncomplementarity was observed in the Politician-of-the-Year Scenario and the Profession Scenario is strong evidence that binary complementarity is not a general property of psychological certainty. However, there are at least two important questions that should be asked. First, did participants understand the MEE nature of the hypotheses? It seems reasonable to conclude that they did. The first sentence of the Politician-of-the-Year Scenario indicated that there are only two candidates remaining for the award, and the Profession Scenario described the exact population—lawyers and engineers—from which the description was drawn. Furthermore, 79% of the participants providing numeric responses exhibited perfect additivity, suggesting that they appreciated the MEE nature of the hypotheses.

The second question asks whether the noncomplementarity that was observed at a group level can be attributed to the responses of only a small portion of participants. If so, the binary complementarity could still be considered a general property of *most* people's psychological certainty. The distribution of verbal overall-certainty scores from the Politician-of-the-Year Scenario suggests that there is little reason to conclude that binary noncomplementarity is restricted to a unique population. For explication purposes, consider an overall-certainty score of 10 as a cut-off point on a verbal scale (although I caution again that an overall score of 10 on the verbal scale does not indicate additivity). In the weak-evidence condition, 19% of the participants had scores that fell below 10, while in the strong-evidence condition, no participants had a score below 10. Other cut-off points yield similar findings. For example, 20% of participants in the weak-evidence condition had scores of 12 or more, compared to 55% in the strong-evidence condition. Although it is impossible to determine a proportion of participants exhibiting complementarity versus noncomplementarity, these figures from the Politician-of-the-Year Scenario suggest that shifts in the means from the overall-certainty scores are not attributable to the responses of only a few participants. Results from the Profession Scenario show a similar trend, but are weak because the overall impact of the evidence-strength manipulation was relatively small for that scenario.

Given the importance of the above two questions for conclusions about binary complementarity, I collected additional data with a third scenario describing a binary case. This Two-Suspect Scenario is printed in Appendix C. The MEE nature of the two events in this scenario was made extremely clear. Also, the evidence-strength manipulation was designed to be stronger than in the Two-Professions Scenario. Participants from the same population as those in Experiment 1 read the scenario after participating in an unrelated experiment in my lab ($N = 151$). All participants provided certainty estimates for both suspects on a verbal scale. The results yielded additional evidence of binary noncomplementarity; overall-certainty scores were significantly higher in the strong-evidence condition ($M = 12.0$, $SD = 1.9$) than in the weak-evidence condition ($M = 10.2$, $SD = 2.5$), $t(149) = 4.91$, $p < .001$, $d = .81$. The distribution of overall-certainty scores in the two conditions makes it clear that the observed effect is not attributable to responses from a small set of participants; 3% of the scores in the strong-evidence condition fell below 10, whereas 27% of the scores in the weak-evidence condition fell below 10.

EXPERIMENT 2

Experiment 1 used a nontraditional verbal measure of certainty to demonstrate binary noncomplementarity. Do responses on this verbal measure reflect the perceptions of certainty that actually mediate people's decisions and behavior? There is good reason to assume that they do. Although numerous studies have compared various aspects of verbal versus numeric expressions of certainty (for a recent review see Budescu & Wallsten, 1995), none have suggested that verbal expressions fail to reflect the perceptions of certainty that drive behavior. In fact, in research involving the same numeric and verbal scales used in Experiment 1, Windschitl and Wells (1996) demonstrated that relative to numeric measures, verbal measures can be better predictors of people's preferences and behavioral intentions under uncertainty. Participants in their third experiment read scenarios with unknown outcomes and provided either verbal or numeric certainty estimates for a focal outcome (e.g., "How likely is it that you actually won a free TV and VCR?"). Later they reread the scenarios and indicated how they would behave in such a situation (e.g., "How many miles would you travel to check into your prize?").[3] Responses on the verbal measures were significantly better predictors of behavioral intentions than were responses on the numeric measure.

Although these findings provide evidence that verbal measures tap into the certainty that drives judgments and decisions under uncertainty, one could argue that, with respect to the issue of complementarity, numeric measures

[3] Some participants provided behavior intentions before providing certainty estimates. The order in which they answered these questions had no significant effect on the results.

provide a more accurate picture of how people typically reason under uncertainty. For example, it could be argued that when faced with a decision based on events in a binary case, people routinely consider both the focal event and its complement before proceeding. In support of this idea, one could note that when deciding whether to buy tickets to see your favorite basketball team (and assuming you want to see the game only if it is a victory), you might consider both the quality of your team and the quality of the other team. However, although it is certainly plausible that many types of decisions prompt a full consideration of the evidence for both possible events, this does not necessitate that all decisions under uncertainty are mediated by perceptions of certainty that conform to binary complementarity. Experiment 2 was designed, in part, to demonstrate that binary noncomplementarity can also be exhibited in decisions. Such a demonstration would indicate that violations of binary complementarity are not restricted to verbal certainty responses. Instead, such violations are properties of the internal perceptions of certainty that drive judgments, decisions, and behaviors.

Participants in Experiment 2 were asked to imagine that they were in a person-perception workshop, in which there were two boxes that contained descriptions of individuals. One box contained descriptions of only politicians and journalists, and the other contained descriptions of only teachers and librarians. Participants saw one description from each box. The description from the politician/journalist box described a person named Sally. Sally's description provided strong evidence that she was a politician but also strong evidence that she was a journalist. The description from the teacher/librarian box described a person named Pat. Pat's description provided weak evidence (virtually no evidence) that he was a teacher but also weak evidence that he was a librarian. The key dependent measure required a choice from participants: "Imagine that the workshop coordinator gave you a five dollar bill and told you to bet the money on one of the following two statements: 1) Sally is a politician, 2) Pat is an elementary school teacher." Half of the participants saw the complementary events in this betting question: "1) Sally is a journalist, 2) Pat is a librarian."

If the subjective certainty guiding participants' decisions conformed to binary complementarity and additivity, then the overall number of people betting their money on knowing Sally's occupation should be roughly equivalent to the number of people betting their money on knowing Pat's occupation. As a specific supporting example, imagine a participant whose perceptions of certainty conform to binary complementarity and who believes there is an 80% chance that Sally is a politician (and 20% chance she is a journalist) and a 70% chance that Pat is a teacher (and 30% chance he is a librarian). If asked to place a bet on "Sally is a politician" or "Pat is a teacher," the participant would place his/her money on "Sally is a politician," but if asked to choose between the complementary events—"Sally is journalist" or "Pat is librarian"—the participant would place his/her money on "Pat is a librarian." This helps illustrate that, because exactly half of the participants in Experiment 2 were asked about the first pair of events and half were asked about the complementary pair, there should be no overall preferences for betting on statements about Sally versus Pat if binary complementarity holds.

The prediction, however, was that binary complementarity would not hold. When people are judging certainty, they will primarily consider the strengths of the focal hypotheses (i.e., those mentioned in the betting questions they read) and neglect the strengths of the complementary hypotheses. The focal hypothesis for Sally—regardless of whether it is "politician" or "journalist"—will always seem to have strong support and the focal hypothesis for Pat will always seem to have weak support. Hence, certainty that Sally is a politician will be greater than certainty that Pat is a teacher, and certainty that Sally is a journalist will also be greater than certainty that Pat is a librarian. This noncomplementarity in subjective certainty will be exhibited in betting decisions: Across the counterbalancing factor, participants will show a general tendency to bet on Sally's occupation rather than Pat's.

This noncomplementarity was expected for only half of the participants in this experiment; a second factor was included to test whether soliciting subjective probability estimates would eliminate or reduce noncomplementarity. Half of the participants were asked to provide subjective probability estimates—that Sally was a politician, Sally was a journalist, Pat was a teacher, and Pat was a librarian—prior to deciding how they would bet. Because this question would force participants to partition their certainty on a fixed scale, their perceptions of certainty would be constrained to follow complementarity, and this complementarity would be observed in their betting decisions.

Method

Participants and design. The participants were 372 students from introductory psychology courses at the University of Iowa. The design was a 2 × 2 × 2 between-subjects factorial. Half of the participants provided betting decisions before providing subjective probability estimates, and half did the reverse. Half of the participants were asked to bet on "Sally is a politician" or "Pat is an elementary school teacher," and half were asked to bet on "Sally is a journalist" or "Pat is a librarian." Finally, half of the participants read questionnaires in which Pat was mentioned before Sally in the main scenario and in each question, and half saw questionnaires in which this order was reversed. As expected, this counterbalancing factor had no significant effects on the results and is therefore excluded from the analyses described in the results section.

Procedure and materials. Participants read questionnaire packets that contained the main scenario, a betting question, a confidence question, and subjective probability questions. All participants read the following scenario (with the last two paragraphs in counterbalanced order).

> Imagine that you and a friend are in a person-perception training workshop . . . One training session involves being able to distinguish between real-life categories of people based on short descriptions prepared by psychologists who interviewed them. There are two boxes full of descriptions.
>
> One box contains descriptions of librarians and elementary school teachers. About half of the

> descriptions are from librarians and half are from elementary school teachers. A description pulled from this box reads: “Pat is 23 and single. His favorite activities include rock climbing, dating, and playing the bass in his rock band. He was engaged to be married twice, but broke off both engagements because he ‘wasn’t ready to settle in yet’.”
> The other box contains descriptions of journalists and politicians. About half of the descriptions are from journalists and half are from politicians. A description pulled from this box reads: “Sally is 37, married, and has 2 children. She has a masters degree in communication studies and worked for 2 years as public relations director for the Windam’s Advocacy Group. She has excellent interpersonal skills, is highly motivated to perform, and shows several positive leadership qualities.”

After reading the scenario, participants first responded either to the betting and confidence questions or to the subjective probability questions. One version of the betting question read:

> Imagine that the workshop coordinator gave you a five dollar bill and told you to bet the money on one of the following two statements: 1) Pat is an elementary school teacher, 2) Sally is a politician. If the statement that you select happens to be correct, you will win another five dollars. If the statement that you select is not correct, you will lose the five dollar bill. On which of the two statements would you bet your money?

Other versions of this question had different response options (Pat is a librarian, Sally is a journalist) and counterbalanced orders (the Sally option was mentioned before the Pat option). The confidence question asked participants to indicate how confident they were in their response on a 7-point scale (1 = *not at all*; 7 = *totally*).

The subjective probability measure solicited estimates for the two occupations of both Sally and Pat. Below are the instructions and the individual questions, which were shown in counterbalanced orders.

> Please answer each of the following questions with a numeric chance estimate between 0% and 100%. For example, a response of 25% would mean that you think there is a one-in-four chance that the statement is true. Feel free to use any number between 0% and 100%.
> What is the chance that Pat is a librarian? ________%
> What is the chance that Pat is an elementary school teacher? ________%
> What is the chance that Sally is a journalist? ________%
> What is the chance that Sally is a politician? ________%

Results and Discussion

The initial question of interest concerns those participants who made their betting decisions before providing subjective probability estimates (i.e., betting-first participants). Specifically, did the betting decisions of those participants exhibit evidence of noncomplementarity? If the perceptions of certainty that mediated their betting decisions were fully complementary, then there should be no overall preferences for betting on Sally’s occupation versus Pat’s. As predicted, however, substantially more of the betting-first participants preferred to bet on Sally’s occupation than on Pat’s, $\chi^2(1, N = 186) = 31.1, p < .001$. The results displayed in Table 2 show that participants who were asked to bet on “Sally is a politician” or “Pat is a teacher” tended to prefer “Sally is a politician” $\chi^2(1, N = 93) = 4.7, p < .05$. Participants who were asked to bet

TABLE 2

Participants' Betting Choices as a Function of Question Order (Betting or Probability First) and Which Occupations Were Asked About in the Betting Question

	% Betting on	
Question order/occupations in the betting question	Sally	Pat
Betting question first		
Sally is politician or Pat is teacher? ($n = 93$)	61.3*	38.7
Sally is journalist or Pat is librarian? ($n = 93$)	79.6*	20.4
Total ($n = 186$)	70.4*	29.6
Probability question first		
Sally is politician or Pat is teacher? ($n = 93$)	47.3	52.7
Sally is journalist or Pat is librarian? ($n = 93$)	67.7*	32.3
Total ($n = 186$)	57.5*	42.5

Note. Asterisks indicate that participants in the respective groups exhibited a statistically significant trend toward choosing Sally's occupation rather than Pat's.

on "Sally is a journalist" or "Pat is a librarian" tended to prefer "Sally is a journalist" $\chi^2(1, N = 93) = 32.5$, $p < .001$. Hence, participants appeared to have had relatively high certainty that Sally is a politician and that Sally is a journalist and relatively low certainty that Pat is a teacher and that Pat is a librarian. This finding provides strong evidence of binary *non*complementarity.

The second question of interest is whether noncomplementarity exhibited in betting decisions was eliminated or reduced when participants first provided subjective probability estimates (the probability-first group). Somewhat surprisingly, there was a significant tendency for probability-first participants to choose to bet on Sally's occupations rather than Pat's (again suggesting noncomplementarity), $\chi^2(1, N = 186) = 4.2$, $p < .05$. However, as predicted, this tendency was substantially reduced in comparison to betting-first participants, $\chi^2(1, N = 372) = 6.72$, $p = .01$. This result suggests that soliciting probability estimates from participants changed the way in which they thought about their uncertainty, and as a result, changed the betting decisions that they made. The explanation offered here is that the probability questions prompted participants to partition their certainty along a fixed scale and required that they attend to both the focal hypothesis and the complementary hypotheses for both Sally and Pat.

The above explanation includes the assumption that, when responding to the probability questions, participants partitioned their certainty on a fixed scale. However, some participants may have failed to realize that the "politician" and "journalist" were described as MEE hypotheses, as were "teacher" and "librarian." Also, some participants might have known that the hypotheses were MEE (not, of course, in these terms), but were unaware of how this fact should influence their probability estimates; personal observations across numerous studies in my laboratory suggest that a small but notable portion of undergraduate students use probability scales in ways that are not remotely consistent with how they are commonly or formally used. Is the same pattern of results observed when participants in these two groups (those who didn't realize the MEE nature of the hypotheses and those who didn't know the implications of MEE hypotheses) are excluded from the analyses?

A second set of analyses were conducted that included only those participants who gave probability estimates that were perfectly additive. That is, their probability estimates for Sally being a journalist and for Sally being a politician summed to exactly 100, and their estimates for Pat being a teacher and Pat being a librarian summed to exactly 100. Clearly, these *perfectly additive* participants understood the MEE nature of the hypotheses and how this fact should constrain probabilities. Nevertheless, as was the case for the full pool of participants, the perfectly additive participants who made a betting decision before estimating probabilities tended to bet on Sally's occupation rather than on Pat's, $\chi^2(1, N = 121) = 9.00, p < .01$ (see Table 3 for the frequencies data). This suggests that, at the time they were making their betting decisions, their perceptions of certainty mediating those decisions did not conform to complementarity. Only when they were asked for subjective probabilities were they forced to partition their certainty, resulting in probability estimates that conformed to additivity and complementarity. Another important finding regarding the perfectly additive participants is that a chi-square test comparing betting choices in the betting-first versus probability-first conditions was significant, $\chi^2(1, N = 261) = 4.91, p < .05$. The perfectly additive participants who estimated probabilities before making a betting decision exhibited no trend for betting on Sally's occupation versus Pat's, $\chi^2(1, N = 140) = 0.0$. Hence, soliciting probability estimates before betting decisions appears to have wiped out any evidence of noncomplementarity in betting decisions.

Analyses of the confidence questions support the same conclusions. Recall

TABLE 3

Betting Choices of Participants Exhibiting Perfect Complementarity in Their Probability Estimates

	% Betting on	
Question order/occupations in the betting question	Sally	Pat
Betting question first		
Sally is politician or Pat is teacher? (n = 62)	54.8	45.2
Sally is journalist or Pat is librarian? (n = 59)	72.9*	27.1
Total (n = 121)	63.6*	36.4
Probability question first		
Sally is politician or Pat is teacher? (n = 76)	44.7	55.3
Sally is journalist or Pat is librarian? (n = 64)	56.3	43.8
Total (n = 140)	50.0	50.0

Note. Asterisks indicate that participants in the respective groups exhibited a statistically significant trend toward choosing Sally's occupation rather than Pat's.

that immediately after making a betting decision, participants indicated confidence in their decision (1 = not confident at all; 7 = totally confident). Confidence responses were recoded onto a −6 to +6 scale. On this recoded scale, a −6 reflected high confidence in a decision to bet on Pat's occupation, a +6 reflected high confidence in a decision to bet on Sally's occupation, and a 0 reflected no confidence at all (i.e., a response of 1 on the actual scale seen by the participants). When all participants are included, a *t* test indicates that the recoded confidence scores were significantly higher for betting-first participants ($M = 1.48$) than for probability-first participants ($M = 0.60$), $t(370) = 2.28$, $p < .05$. When only perfectly additive participants are included, the difference between betting-first ($M = 0.92$) and probability-first participants ($M = 0.05$) was nearly significant $t(370) = 1.83$, $p = .07$. Hence, soliciting probability estimates from participants before they made betting decisions tended to lower their confidence in betting on Sally's occupation (or increase their confidence in betting on Pat's occupation).

The data from the probability questions are summarized in Table 4 (for the full pool of participants) and Table 5 (for the perfectly additive participants). Not surprisingly, when nonadditive participants are included in the analysis, probability estimates for Sally's occupations are greater than those for Pat's occupations. For perfectly additive participants, estimates for Sally's and Pat's occupations are, by necessity, equivalent. There was no evidence that participants' probability estimates were significantly influenced by having first made betting decisions.

In summary, the results for the betting-first group provide evidence of binary *non*complementarity in people's decisions. More specifically (and perhaps more accurately), the results demonstrate that perceptions of certainty mediating people's decisions did not conform to the constraints of binary complementarity. When asked to place their money on one of two independent events (e.g., that Sally is a politician or Pat is a teacher), they tended to choose the event that

TABLE 4

Probability Estimates for the Four Occupations (Includes All Participants)

Question order/occupations in betting question	Probabilities for Sally		Probabilities for Pat	
	Politician	Journalist	Teacher	Librarian
Betting question first				
Sally is politician or Pat is teacher?	54.6 (20.2)	44.5 (21.5)	50.4 (23.9)	33.8 (20.4)
Sally is journalist or Pat is librarian?	48.6 (21.4)	51.7 (23.0)	47.7 (20.0)	38.2 (18.7)
Total (*n* = 186)	51.6 (21.0)	48.1 (22.5)	49.0 (22.0)	36.0 (19.6)
Probability question first				
Sally is politician or Pat is teacher?	50.8 (21.6)	48.3 (22.1)	56.8 (21.3)	36.2 (19.8)
Sally is journalist or Pat is librarian?	54.6 (22.1)	46.7 (21.7)	53.3 (22.0)	35.7 (20.1)
Total (*n* = 186)	52.7 (21.9)	47.5 (21.8)	55.0 (21.6)	36.0 (20.0)

Note. Regardless of what occupations were listed in the betting question, all participants gave probability estimates for all four occupations.

TABLE 5

Probability Estimates for the Four Occupations from Perfectly Additive Participants Only

Question order/occupations in betting question	Probabilities for Sally		Probabilities for Pat	
	Politician	Journalist	Teacher	Librarian
Betting question first				
Sally is politician or Pat is teacher?	54.5 (19.4)	45.1 (20.0)	58.2 (18.1)	41.8 (18.1)
Sally is journalist or Pat is librarian?	49.2 (19.1)	50.9 (19.1)	55.9 (12.9)	44.1 (12.9)
Total (*n* = 186)	51.9 (19.4)	47.9 (19.7)	57.1 (15.7)	42.9 (15.7)
Probability question first				
Sally is politician or Pat is teacher?	49.9 (21.3)	50.1 (21.3)	61.0 (17.5)	39.1 (17.5)
Sally is journalist or Pat is librarian?	55.2 (21.3)	44.8 (21.3)	58.4 (19.5)	41.6 (19.5)
Total (*n* = 186)	52.4 (21.4)	47.6 (21.4)	59.8 (18.4)	40.2 (18.4)

Note. Perfectly additive participants are those whose probability responses summed to exactly 100 for each of the two MEE event pairs in Experiment 2. Hence, the mean responses for those pairs some to 100 in this table, excluding rounding error.

had the most support, which was always the occupation mentioned for Sally, even though the complement to this event also received strong support. Hence, when judging certainty prior to making a decision, participants appear to have overweighted the evidence for the focal hypotheses (those explicitly mentioned in the betting question) and underweighted the evidence for the complementary hypotheses. Had the evidence for the focal and complementary hypotheses received equal weight, there would have been no overall preferences for betting on Sally's occupation rather than Pat's. The perfectly additive participants in the probability-first group showed no such preference; the binary complementarity that was prompted by the probability questions influenced the subsequent betting decisions.

GENERAL DISCUSSION

The two assertions that were made in the introduction received strong support from the data. First, full binary complementarity is not a general property of people's certainty. Binary noncomplementarity was observed on both verbal and numeric measures in Experiment 1, and the pattern of betting decisions of participants in Experiment 2 revealed noncomplementarity in internal perceptions of certainty. Second, numeric measures of subjective probability overestimate the degree to which people's perceptions of certainty conform to complementarity. Whereas the noncomplementarity detected with numeric measures in Experiment 1 was minimal, the noncomplementarity detected with verbal measures of certainty was quite robust. Experiment 2 demonstrated that the noncomplementarity observed on the betting measure can be significantly moderated by first soliciting numeric probability estimates from participants.

Sources of Noncomplementarity

Why was noncomplementarity observed in the present research? Researchers have taken at least four distinct, but somewhat overlapping approaches to explaining noncomplementarity. First, some researchers have discussed the idea of independent confidence (Van Wallendael & Hastie, 1990) or a nondistributional conception of probability (Teigen, 1983). These views suggest that people represent certainty in MEE hypotheses not as a set amount, but as individual possibilities for which certainty in one can vary independently of certainty in the others. McKenzie's (1998, 1999) work extended the notion of independent confidence by demonstrating how noncontrastive learning can encourage the development of independent confidences in two MEE hypotheses. As described earlier, noncontrastive learners—who underwent a training phase that taught them whether each of a series of symptoms was or was not diagnostic of one disease and was or was not diagnostic of a second disease—later gave nonadditive probability estimates for patients' chances of having the MEE diseases.

A second, related proposal for explaining noncomplementarity focuses on biased hypothesis testing. Sanbonmatsu, Posavac, and Stasney (1997) described how the confirmation bias can influence probability judgment. Participants in their research, who were asked to assess the likelihood that one of four candidates was hired for a position, tended to focus on evidence relevant to the focal candidate while ignoring evidence relevant to the alternative candidates. Consequently, participants' probability judgments were inflated and exhibited nonadditivity when the evidence for all candidates was favorable.

A third proposal for explaining noncomplementarity comes from Tversky and Koehler's (1994) support theory, which suggests that noncomplementarity occurs (only in nonbinary cases) because alternatives to a focal hypothesis tend to be left unpacked as implicit disjunctions. For example, when asked to judge the probability that Indiana University would win the Big 10 Conference title, the typical respondent would leave the alternative hypothesis unpacked ("all other Big 10 teams") and would likely underestimate the evidence for the components of that alternative hypothesis, resulting in an inflated probability estimate for the focal hypothesis.

Finally, as part of an extension to support theory called asymmetric support theory, Brenner and Rottenstreich (1999) proposed a type of nonadditivity that can result (even in binary cases) when the relevant hypotheses are somewhat malleable. For example, the theory suggests that for probability judgments based on assessments of category size (e.g., What is the probability that a randomly selected UCLA alumnus is a salesperson rather than a social worker?), the representation of a hypothesis tends to be tighter (i.e., more strictly or narrowly construed) when it is focal rather than an alternative. Assuming that less evidence will seem to apply to a tighter rather than looser representation, the probability responses for complimentary versions of such questions will tend to sum to less than 1.0.

All four of the above proposals are backed by strong empirical evidence, and all four are likely to be valid descriptions for how noncomplementarity can arise under certain conditions. However, for explaining the noncomplementarity observed in the present research, the proposal from support theory can be ruled out; failure to unpack the alternative hypothesis into components can explain nonadditivity in nonbinary cases but not in binary ones. Also, the proposal from asymmetric support theory can be ruled out because there is no apparent malleability in the events described in scenarios that showed strong evidence of binary complementarity (i.e., the Politician-of-the-Year Scenario and Two-Suspect Scenario).

The account I propose for why noncomplementarity was observed in the present research is more closely related to ideas of McKenzie (1998) and Sanbonmatsu et al. (1997). Specifically, when people are assessing certainty, evidence that is most directly related to a hypothesis will have more impact than evidence that is one step removed from that hypothesis. In the scenarios used here, evidence specific to the focal hypothesis was directly relevant to determining its likelihood (regardless of what the alternative hypothesis was), but the evidence specific to the alternative hypothesis was relevant to determining the likelihood of the focal hypothesis only because the two hypotheses had been linked together as a MEE set. This is somewhat analogous to the experience of participants in McKenzie's noncontrasted learning group, who underwent a training phase that taught them whether each of a series of symptoms was or was not diagnostic of one disease and was or was not diagnostic of a second disease. By normative standards, a piece of evidence that was not *directly* diagnostic for one disease could nevertheless be important for assessing the probability of that disease because the evidence was diagnostic of the second disease (linked in a MEE set). However, even if participants in McKenzie's research and the present research were aware that evidence for both hypotheses in a MEE set was relevant to judging the focal hypothesis, evidence specific to the alternative hypothesis may have had less impact because its connection to the focal hypothesis was less direct and dependent upon an arbitrary link between the two hypotheses. People's hypothesis testing strategies would likely be influenced in a related manner. That is, people would be biased toward assessing evidence that has a direct link to the focal hypothesis rather than evidence that is linked only on an arbitrary basis.

This proposal is consistent with recent research on how the familiarity of events can have a biasing influence on judgments of relative likelihood (Fox & Levav, 1998). This research indicates that people sometimes view familiar events and their familiar complements as more likely than unfamiliar events and their unfamiliar complements. If asked who will *win* an upcoming game between University of Washington (familiar) and Harper State (unfamiliar), a respondent might say University of Washington; but if asked who will *lose* the same game, the same respondent might also say University of Washington (see Shafir, 1993, for related work). Fox and Levav suggest that recruiting support for a familiar event or for its familiar complement is easier than recruiting support for an unfamiliar event or its complement. In situations where support for focal events receives more weight than support for nonfocal events, judgments of certainty will tend, overall, to favor familiar focal events over unfamiliar focal events. Fox and Levav argued that this differential weighting of support for focal and nonfocal events will be especially prominent for judgments of relative likelihood (Which is more likely?), but less prominent for judgments of absolute likelihood (How likely is X?). In their studies, the effects of familiarity bias were robust for relative judgments but not for absolute judgments. Although the differential weighting of support for focal and nonfocal hypotheses may in fact be more influential for judgments of relative likelihood than for judgments of absolute likelihood, the noncomplementarity observed in Experiment 1 of the present research indicates that this differential weighting of support can be quite robust for absolute judgments as well.

Sources of Complementarity

A different approach to understanding binary noncomplementarity focuses more on why and how perceptions of certainty exhibit complementarity rather than on why such perceptions fail to conform to full binary complementarity. Although the present experiments provided demonstrations in which people were not exhibiting full binary complementarity, this does not mean that complementarity was completely absent. As seems to be typical in most everyday situations involving two MEE hypotheses, certainty in one of the hypotheses in Experiment 1 decreased as certainty in the other increased. What mechanisms help achieve complementarity in everyday judgment and decision making?

McKenzie (1998) suggested that participants in his contrasted learning conditions developed dependent confidence, where confidence in one hypothesis is the polar opposite of confidence in the other. Recall that in his contrasted learning conditions, participants were essentially trained to view individual disease symptoms as either diagnostic of one disease or another disease (in a MEE pair). Hence, in a test phase, if most of a hypothetical patient's symptoms fit the first disease, confidence in the second disease would necessarily be low. In everyday environments, a person might have repeated experiences distinguishing between the same two MEE hypothesis. In such cases, the person might learn to *simultaneously* view a piece of evidence as supportive of one hypothesis and contrary to the other (McKenzie, 1998).

Although McKenzie's (1998) idea of dependent confidence describes one mechanism for achieving complementarity, the partial complementarity observed in the present experiments requires a different explanation because the hypothesis pairs were novel to the participants. It seems reasonable to assume that the complementarity was achieved through a weighing process in which the support for the focal hypothesis was compared to the support for the alternative hypothesis. This weighing process is the same mechanism that support theory proposes to account for the binary complementarity of subjective probability estimates (Tversky & Koehler, 1994). The present work suggests that for internal assessments of certainty, a weighing process is operating but is biased in the sense that evidence for a focal hypothesis has more impact than evidence for the alternative hypothesis.

Finally, the present experiments suggest an additional mechanism by which complementarity (especially perfect complementarity) might be achieved. Specifically, in a response-generation phase, a person might recognize that a given set of responses does not conform to the rules of probability. For example, they might recognize that their responses should not equal 134% or that they should not respond “very likely” to two MEE hypotheses. The person might then revise their responses accordingly. In fact, the results of Experiment 2 suggest that revisions prompted by the response-generation phase can lead to changes in internal perceptions of certainty as well. Specifically, the betting decisions in Experiment 2 were consistent with binary complementarity only after participants generated probability estimates, suggesting that the response-generation stage influenced the perceptions of certainty that mediated their betting decisions.

Verbal versus Numeric Measures of Certainty

The finding that verbal measures were less likely to elicit fully complementary responses than were numeric probability measures might appear to be at odds with related research that focused on the calibration and coherence of verbal and numeric certainty judgments (Wallsten, Budescu, & Zwick, 1993). Participants in that research provided both verbal and numeric certainty judgments for a large number of general-knowledge items (300). Each general-knowledge item was cast in two complementary ways (“The Monroe Doctrine was proclaimed before the Republican Party was founded” and “The Republican Party was founded before the Monroe Doctrine was proclaimed”), and these two forms were seen in separate testing sessions. Participants verbal responses were quantified using two separate techniques (see Wallsten et al. for details), and the additivity of participants’ quantified responses to the two complementary claims was assessed. In analyses across the 300 items, both numeric and verbal responses (regardless of the technique used for quantification) appeared to exhibit nearly perfect additivity, and no differences in additivity (or complementarity) were observed between the two response modes.

Despite the appearance of a contradiction between those results and the results of Experiment 1, three key differences between the studies should be noted. First, the items used by Wallsten et al. (1993) were quite different from the scenarios used in the present research. Given the way in which complementary claims were constructed for Wallsten et al.’s items, it seems unlikely that participants could have developed independent representations of confidence (as opposed to dependent confidence) for the two claims. Any piece of evidence or knowledge that a respondent might recall in favor of one claim would directly and simultaneously contradict the complementary claim. The scenarios in the present research were constructed such that evidence for one hypothesis was quite distinct from evidence for the other hypothesis(es), which would be more likely to encourage independent representations of confidence. Second, Experiment 1 involved a manipulation of support (strong for both hypotheses or weak for both hypotheses) that led to the detection of noncomplementarity. Support was not manipulated in the Wallsten et al. study, so conditions were not especially favorable for finding evidence of noncomplementarity. Third, Wallsten et al. assessed additivity by collapsing across items. While this strategy allows for an assessment of overall tendencies of verbal and numeric responses to be additive, it does not focus on the additivity of responses to individual items, as was done in the present research. In summary, given the nature of the differences between Wallsten et al.'s experiment and Experiment 1, there is no reason to view their results as contradictory. Experiment 1 used conditions that were favorable for the detection of noncomplementarity and provided a clear demonstration in which, relative to the verbal measure, the numeric measure underestimated the degree of noncomplementarity in perceptions of certainty.

Do the present findings suggest that researchers should choose verbal measures over numeric ones? Recent research has illustrated some important advantages (yet also disadvantages) that verbal measures have for assessing people's perceptions of certainty (Windschitl & Weber, 1999; Windschitl & Wells, 1996, 1998). The fact that numeric measures can overestimate the complementarity of internal perceptions of certainty suggests a possible additional advantage for verbal measures. Imagine a respondent who learns about two MEE hypotheses that both have high and roughly equivalent levels of support. If asked to give verbal certainty estimates, the person might say "fairly likely" and "very likely" for the two hypotheses; if asked to give numeric certainty estimates, the person might regress his/her estimates to "45%" and "55%," respectively. For a researcher interested in people's internal perceptions of certainty given sets of evidence, the verbal measures might prove more useful. Responses like "45% and "55%" would simply underestimate the certainty that the person has in the hypotheses and could consequently underestimate the confidence and likelihood with which the person would make a decision that is based on one of the two hypothesized events occurring. At a more general level, verbal measures may provide researchers with a more valid picture of how people typically consider and weigh the evidence for focal and nonfocal hypotheses across different types of situations.

It is important to note, however, that there are many conditions in which the numeric measure's enhanced sensitivity to complementarity is not a liability but rather neutral or advantageous. Three examples follow. First, there are some real-world domains in which people scale their own perceptions of certainty in terms of numeric probabilities, and there are types of people who typically consider uncertainty in numeric terms. Also, some types of people, because of their training or career, might almost always ensure that their judgments exhibit appropriate complementarity. For such cases, numeric measures would not overestimate the complementarity of internal perceptions of certainty. Second, when a researcher is interested in how accurate a participant's judgments can be, maximizing participants' concerns with complementarity would likely be beneficial rather than problematic. Third, in situations where the evidence for both the focal and alternative events typically receive equal weight, there would likely be little differences in the degree of complementarity observed with verbal versus numeric measures.

This research was not designed to settle questions regarding the advantages and disadvantages of various certainty measures. Nevertheless, this research is a reminder that the precision and ease-of-use that are achieved with traditional numeric subjective probability measures need to be weighed against some important advantages found in less common nonnumeric measures of certainty (Windschitl & Wells, 1996). The researcher's choice of what method should be used to measure certainty is consequential not only for the findings of an individual study but also for general conclusions about how people make judgments and decisions under uncertainty.

APPENDIX A

Verbal and Numeric Certainty Scales Used in Experiment 1

________ certain
________ extremely likely
________ quite likely
________ fairly likely
________ slightly likely
________ as likely as unlikely
________ slightly unlikely
________ fairly unlikely
________ quite unlikely
________ extremely unlikely
________ impossible

________ 100%
________ 90%
________ 80%
________ 70%
________ 60%
________ 50%
________ 40%
________ 30%
________ 20%
________ 10%
________ 0%

APPENDIX B

Below are the weak- and strong-evidence versions of the scenarios used in Experiment 1.

Politician-of-the-Year Scenario, Weak-Evidence Version

After a long selection process, there are two candidates remaining for Alabama's Republican of the Year Award. Rebecca Sharp is a divorced 36-year-old. She was instrumental in running the campaigns of several Republicans that were recently elected into various offices. Although she is popular among some young Republicans, she is disliked by many of the most powerful male and older Republicans. Stacy Rakan headed an organization called Republican Women for Choice that raised large amounts of money for the campaigns of four female Republicans. She is supported by some women's groups, but conservative Republicans are upset with her for her attempts to keep abortion legal.

How likely is it that Rebecca Sharp will win the Alabama Republican of the Year Award?

How likely is it that Stacy Rakan will win the Alabama Republican of the Year Award?

Politician-of-the-Year Scenario, Strong-Evidence Version

After a long selection process, there are two highly-praised candidates remaining for Alabama's Democrat of the Year Award. Rebecca Sharp is 36 years old, divorced, and a mother of two children. She was instrumental in running the campaigns of several Democrats that were recently elected into various offices. Several groups have encouraged her to run for the Senate. Stacy Rakan is a married, 46-year-old mother of five. She headed an organization called Democratic Women that raised large amounts of money for the campaigns of four female Democrats. She has been encouraged to run for a House of Representatives Seat.

How likely is it that Rebecca Sharp will win the Alabama Democrat of the Year Award?

How likely is it that Stacy Rakan will win the Alabama Democrat of the Year Award?

Profession Scenario, Weak-Evidence Version

Several psychologists interviewed a group of people. The group included 30% engineers and 70% lawyers. The psychologists prepared a brief summary of their impressions of each interviewee. The following description was drawn randomly from the set of descriptions:

Mark is 35 years old, married, and has five children. He should fit well in his field and be liked by his colleagues. He enjoys painting, drawing, music, bicycling, and movies. Mark is quite extroverted and works well in groups.

How likely is it that Mark is an engineer?

How likely is it that Mark is a lawyer?

Profession Scenario, Strong-Evidence Version

Several psychologists interviewed a group of people. The group included 30% engineers and 70% lawyers. The psychologists prepared a brief summary of their impressions of each interviewee. The following description was drawn randomly from the set of descriptions:

Mark is 35 years old, married, and has one child. He likes orderly systems where each item finds its appropriate place. Mark is argumentative, ambitious, and highly articulate in his oral and written expression. His analytic and spatial reasoning abilities are superb. He enjoys reading, puzzles, history, and photography. He also spends time constructing model airplanes and ships.

How likely is it that Mark is an engineer?

How likely is it that Mark is a lawyer?

Senate-Race Scenario, Weak-Evidence Version

Only three candidates are vying for a U.S. Senate seat from Oregon. Gregory Timson is 65 years old and is known as “the owl” by his constituents and his fellow state politicians. He has little support from young voters.

Zackary Brown is a single 36-year-old. He started in politics at age 32 after he had brought a large business from near bankruptcy to financial success. Although he is considered “money smart,” many people distrust his character and some of his business dealings.

Anthony Jenkins is 40 years old, married, and a father of five children. He is known for his ability to rally support for special causes and his innovative ideas. He is, however, thought to be ineffective in implementing and guiding projects, and he has been criticized for not being able to control his staff.

How likely is it that Gregory Timson will win the Senate seat?

How likely is it that Zackary Brown will win the Senate seat?

How likely is it that Anthony Jenkins will win the Senate seat?

Senate-Race Scenario, Strong-Evidence Version

Only three candidates are vying for a U.S. Senate seat from Oregon. Gregory Timson is 59 years old and is known as a “wise owl” by his constituents and his fellow state politicians. He has served in Oregon state politics since 1965 and has a number of long-time supporters who feel well represented by him. Throughout his political career he has sponsored numerous landmark bills that nearly all Oregonians feel have bettered the state.

Zackary Brown is a single 36-year-old who is known for his hard work and skill at forging compromises on difficult issues. He started in politics at age 32 after he had brought a large business from near bankruptcy to financial success. Many Oregonians like his fresh perspectives and are excited to get some “new blood” into the Senate.

Anthony Jenkins is 40 years old, married, and a father of five children. He has worked very hard at reducing the growing crime rate in Oregon. He also has created several programs aimed at helping high school and college graduates find employment opportunities. Jenkins is known for his ability to rally support for special causes and his innovative ideas for balancing the state and federal budgets.

How likely is it that Gregory Timson will win the Senate seat?

How likely is it that Zackary Brown will win the Senate seat?

How likely is it that Anthony Jenkins will win the Senate seat?

Four-Suspect Scenario, Weak-Evidence Version

On the night of February 26, 1994, chemicals from a medical lab in Nebraska were stolen. Only four lab employees know the necessary security codes to enter the lab. The only evidence to suggest which of the employees might have stolen the chemicals is testimony from a witness who got a glimpse of the thief exiting the lab with the chemical at 3:30 A.M. that night. The witness tentatively described the person as “a medium-height woman with short hair.”

The four lab employees are as follows: Tony Martinez is a tall man with long brown hair. Mary Wilson is a medium-tall woman with long hair. Beth Battle is a tall woman with black shoulder-length hair. Pamela Bauman is also a tall woman with black shoulder-length hair.

How likely is it that Tony Martinez was the thief?

How likely is it that Mary Wilson was the thief?

How likely is it that Beth Battle was the thief?

How likely is it that Pamela Bauman was the thief?

Four-Suspect Scenario, Strong-Evidence Version

On the night of February 26, 1994, chemicals from a medical lab in Nebraska were stolen. Only four lab employees know the necessary security codes to enter the lab. The only evidence to suggest which of the employees might have stolen the chemicals is testimony from a witness who got a glimpse of the thief exiting the lab with the chemical at 3:30 A.M. that night. The witness described the person as “a tall woman with black hair cut above the shoulders.”

The four lab employees are as follows: Tony Martinez is a medium-height man with long brown hair. Mary Wilson is a medium-tall woman with short hair that is very dark. Beth Battle is a tall woman with black shoulder-length hair. Pamela Bauman is also a tall woman with black shoulder-length hair.

How likely is it that Tony Martinez was the thief?

How likely is it that Mary Wilson was the thief?

How likely is it that Beth Battle was the thief?

How likely is it that Pamela Bauman was the thief?

APPENDIX C

Mean Certainty Responses for Individual Hypotheses in the Four Scenarios of Experiment 1

Scenario/version/hypothesis	Numeric scale *M* (*SD*)	Verbal scale *M* (*SD*)
Politician of the Year		
Weak-evidence		
Sharp	5.0 (1.4)	5.7 (2.0)
Rakan	4.9 (1.4)	4.5 (1.8)
Strong-evidence		
Sharp	4.8 (1.4)	5.7 (1.9)
Rakan	5.7 (1.5)	6.7 (1.9)
Profession Scenario		
Weak-evidence		
Engineer	4.8 (2.1)	5.7 (2.5)
Lawyer	5.2 (2.5)	5.6 (2.3)
Strong-evidence		
Engineer	4.4 (2.3)	5.4 (2.5)
Lawyer	6.1 (2.1)	6.6 (2.0)
Senate-Race Scenario		
Weak-evidence		
Timson	4.5 (2.0)	6.1 (2.1)
Brown	3.6 (1.5)	5.3 (2.0)
Jenkins	3.5 (1.5)	5.6 (1.8)
Strong-evidence		
Timson	4.7 (1.9)	6.8 (1.8)
Brown	4.0 (1.8)	5.9 (1.9)
Jenkins	3.9 (1.9)	6.5 (1.7)
Four-Suspect Scenario		
Weak-evidence		
Martinez	1.5 (1.7)	2.4 (1.8)
Wilson	4.2 (2.2)	5.8 (2.2)
Battle	3.7 (1.9)	5.6 (2.0)
Bauman	3.7 (1.9)	5.5 (1.9)
Strong-evidence		
Martinez	1.8 (1.9)	2.7 (2.1)
Wilson	2.9 (2.1)	5.1 (2.3)
Battle	4.2 (2.0)	6.9 (1.7)
Bauman	4.2 (2.0)	6.9 (1.8)

APPENDIX D

Below are the weak-evidence and strong-evidence versions of the Two-Suspect Scenario that was used as a follow-up for Experiment 1.

Two-Suspect Scenario, Weak-Evidence Version

On the night of June 24, 1996, chemicals from a government lab in Utah were stolen. A witness got a glimpse of the thief exiting the lab with the chemical at 11:00 P.M. that night. The entrance to the lab was controlled by a high-tech computer that checks a person's hand print before allowing entrance into the lab. The computer was programmed to allow only two people into the lab, employees Sandy Young and Jennette Kahil. The integrity of the computer system was verified by investigators, and hence, the suspect list contained only Young and Kahil.

The witness tentatively described the culprit as "a medium-height woman with short, blond hair." Young, who is 5 foot 1 inches and has medium length reddish-blond hair, claimed she was asleep at the time of the robbery—a story confirmed by her husband. Kahil, who is 5 foot 10 inches and has light brown hair, claimed she was returning from a personal trip to Colorado at the time of the theft. Billing records from a gas station indicate that she was in Colorado on the night of the theft and that it would have been very difficult, but not completely impossible, for her to have been at the lab at the time of the theft. For many reasons, investigators ruled out the possibility that the two women worked together.

Given the evidence, how likely do you think it is that Young was the culprit?

Given the evidence, how likely do you think it is that Kahil was the culprit

Two-Suspect Scenario, Strong-Evidence Version

On the night of June 24, 1996, chemicals from a government lab in Utah were stolen. A witness got a glimpse of the thief exiting the lab with the chemical at 11:00 P.M. that night. The entrance to the lab was controlled by a high-tech computer that checks a person's hand print before allowing entrance into the lab. The computer was programmed to allow only two people into the lab, employees Sandy Young and Jennette Kahil. The integrity of the computer system was verified by investigators, and hence, the suspect list contained only Young and Kahil.

The witness tentatively described the culprit as "a medium-height woman with short, blond hair." Young, who is 5 feet 5 inches tall and has short reddish-blond hair, claimed she was asleep at the time of the robbery but had no one to confirm her story. An envelope containing $5000 in cash was found at her apartment on the day after the theft, but she claimed she had been saving that money for years. Kahil, who is 5 foot 6 inches and has blond hair, claimed she was returning from a personal trip to Colorado at the time of the theft. Like for Young, no one could confirm Kahil's story. Kahil had recently been disciplined by her supervisors for failure to complete some recent projects she was working on, and she had threatened to discotinue all of her projects because she felt she was being paid too little. For many reasons, investigators ruled out the possibility that the two women worked together.

Given the evidence, how likely do you think it is that Young was the culprit?

Given the evidence, how likely do you think it is that Kahil was the culprit?

REFERENCES

Brenner, L., & Rottenstreich, Y. (1999). *Asymmetric support theory: Focus-dependence and context-independence in likelihood judgment*. Unpublished manuscript.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. Busemeyer, R. Hastie, & D. Medin (Vol. Eds.), *The psychology of learning and motivation: Decision making from a cognitive perspective* Vol. 32. (pp. 275–318). New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillside, NJ: Erlbaum.

Fox, C. R., & Levav, J. (1998, November). *Familiarity bias in relative likelihood judgment*. Paper presented at the Meeting of the Society for Judgment and Decision Making, Dallas, TX.

Kirkpatrick, L. A., & Epstein, S. (1992). Cognitive-experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology*, **63,** 534–544.

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 and Cognition*, **27,** 152–165

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.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review*, **104,** 406–415.

Sanbonmatsu, D. M., Posavac, S. S., & Stasney, R. (1997). The subjective beliefs underlying probability overestimation. *Journal of Experimental Social Psychology*, **33,** 276–295.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, **21,** 546–556.

Teigen, K. H. (1974). Subjective sampling distributions and the additivity of estimates. *Scandinavian Journal of Psychology*, **15,** 50–55.

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

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, **102,** 269–283.

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 footsteps of Sherlock Holmes: Cognitive representations of hypothesis testing. *Memory and Cognition*, **18,** 240–250.

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

Windschitl, P. D., & Martin, R. (1999). *Social comparisons influence people's interpretations of objective vulnerability information*. Manuscript in preparation.

Windschitl, P. D., & Weber, E. U. (1999). The interpretation of "likely" depends on the context, but "70%" is 70%—right?: The influence of associative processes on perceived certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **25,** 1514–1533.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, **2,** 343–364.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology*, **75,** 1411–1423.

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

Zimmer, A. C. (1983). Verbal vs. numeric processing of subjective probabilities. In R. Scholz (Ed.), *Decision making under uncertainty* (pp. 159–182). Amsterdam: North–Holland.

Received March 29, 1999