ELSEVIER

Acta Psychologica 113 (2003) 263–282

acta psychologica

www.elsevier.com/locate/actpsy

# Hypothesis generation, probability judgment, and individual differences in working memory capacity

Michael R.P. Dougherty *, Jennifer E. Hunter

*Department of Psychology, University of Maryland, College Park, MD 20742, USA*

Received 25 March 2002; accepted 24 February 2003

**Abstract**

This research examined the role of working memory (WM) in probability judgment and hypothesis generation using a simulated task that involved estimating the likelihood that particular menu items would be ordered by customers at a dinner. Five main findings were observed. First, judgments of the likelihood of individual items were made relative to alternatives retrieved from long-term memory. Second, the number of alternatives retrieved was positively correlated with a measure of WM-capacity (the operation-span task). Third, participants' probability judgments were subadditive (summing to well over 100%). Fourth, the degree to which participants' judgments were subadditive was affected by the number and strength of the alternatives retrieved from long-term memory. Fifth, the degree to which participants were subadditive was negatively correlated with WM-capacity. The results suggest that individual differences in WM-capacity are fundamental to hypothesis generation and probability judgment.


*PsycINFO classification:* 2343; 3040
*Keywords:* Probability judgment; Frequency judgment; Hypothesis generation; Alternative generation; Working memory; Support theory

* Corresponding author.
*E-mail address:* mdougherty@psyc.umd.edu (M.R.P. Dougherty).

0001-6918/03/$ - see front matter 
doi:10.1016/S0001-6918(03)00033-7

## 1. Introduction

Hypothesis generation is an important component of many real-world tasks: Physicians generate disease hypotheses prior to issuing diagnoses (Barrows, Norman, Neufeld, & Feightner, 1982; Elstein, Shulman, & Sprafka, 1978), mechanics generate hypotheses for auto failures (Mehle, 1982), and auditors generate hypotheses regarding accounting errors (Libby, 1985). The hypothesis generation process has obvious, and often profound implications for a variety of decision tasks, affecting physicians' treatment decisions, auditors' decisions to invest time and money pursuing possible sources of errors in accounting records, and mechanics' course of action in repairing automobiles. In all of these cases, the ultimate course of action rests heavily on the decision maker's ability to generate the correct cause underlying an observed pattern of symptoms. As Barrows et al. (1982) has shown, one is unlikely to consider causes or hypotheses that are not initially generated. Thus, the hypothesis generation process is critical for accurate judgment.

Intimately tied to the hypothesis generation process is the process of evaluating the probability or likelihood of specific hypotheses—the process of hypothesis evaluation. When one evaluates the likelihood of a particular hypothesis, it presumably is evaluated relative to other alternative or rival hypotheses. These rival hypotheses must be retrieved or generated from memory (Dougherty, Gettys, & Thomas, 1997; Gettys & Fisher, 1979; Koehler, 1994).

The assumption that the judged probability of a focal hypothesis is made by comparing it to an alternative or alternatives is embodied in several recent theories of probability judgment. For example, Tversky and Koehler's (1994) support theory assumes that the judged probability of a focal hypothesis is assessed by comparing the strength of the evidence for the focal hypothesis to the strength of the evidence of alternative hypotheses. Similarly, Windschitl and Wells (1998) proposed the comparison heuristic, whereby probability judgments were assumed to be based on a comparison of the strength of the focal hypothesis with the single strongest alternative (see also Windschitl & Young, 2001; Windschitl, Young, & Jensen, 2002). Finally, Dougherty (2001; Dougherty, Gettys, & Ogden, 1999) has argued that probability judgments are made by comparing the memory strength of the focal hypothesis with the combined memory strengths of explicitly generated alternatives.

Of both theoretical and practical interest is the question regarding the factors that affect hypothesis generation and hypothesis evaluation. Whereas there has been considerable research investigating hypothesis *evaluation* in both the cognitive and social literatures (Koehler, 1991; Sanbonmatsu, Posavac, Kardes, & Mantel, 1998), relatively little research has examined the process of hypothesis *generation*. Hypothesis generation is the logical precursor to hypothesis evaluation, since the outcome of the generation process determines which hypotheses ultimately are evaluated.

Although relatively little research has been done on hypothesis generation, what has been done has yielded a consistent pattern of results. Of particular importance is the finding that participants generate only a fraction of the total number of plausible alternatives (Fisher, Gettys, Manning, Mehle, & Baca, 1983; Gettys, Pliske, Man-

ning, & Casey, 1987; Mehle, Gettys, Manning, Baca, & Fisher, 1981). For example, in one study Mehle (1982) found that auto mechanics were deficient at generating hypotheses for why a car would not start—generating only around 4–6 total hypotheses. Moreover, participants were overconfident in the sets of hypotheses they did generate, believing that they were more complete than they actually were. More recent research has shown that participants tend to generate alternatives that are highly likely, ignoring hypotheses that are implausible or unlikely (Dougherty et al., 1997; Weber, Böckenholt, Hilton, & Wallace, 1993), and that the judged probability of the focal hypothesis decreases as the number of alternatives generated increases (Dougherty et al., 1997).

The above studies indicate that hypothesis generation is an important component of probability assessment, as the number and strength of alternative hypotheses affects judged probability. However, while these studies have furthered our understanding of hypothesis generation and probability assessment, they have not explored the underlying memory processes. The purpose of the present research was to provide an initial examination of the relationships among probability judgment, hypothesis generation, and working memory processes. In particular, we tested the idea that individual differences in working memory (WM) capacity are fundamental to probability judgment by limiting the number of alternative hypotheses that can be maintained in the focus of attention. As prior research has shown, judged probability tends to decrease as the number of alternatives explicitly considered increases (Dougherty et al., 1997). We propose that one important constraint on the number of alternatives people consider is WM-capacity.

There are at least two properties of working memory functioning that seem relevant to hypothesis generation. One property involves the maintenance of task-*relevant* information in the focus of attention. Considerable research has shown that the number of items one can maintain in the focus of attention is quite small ($4 \pm 1$, see Cowan, 2001) and that it varies among individuals (Engle, Kane, & Tuholski, 1999; Engle, Tuholski, Laughlin, & Conway, 1999). The second property involves the inhibition or suppression of task-*irrelevant* information (Hasher & Zacks, 1988). The extent to which people fail to inhibit task-irrelevant information from entering the focus of attention affects the amount of relevant information that can be maintained. As Engle, Tuholski, Laughlin, and Conway (1999) argued, behavior is more likely to be influenced by task-irrelevant information when it cannot be inhibited from entering the focus of attention (see also Kane, Bleckley, Conway, & Engle, 2001; Rosen & Engle, 1997).

The ability to maintain task-relevant information in the focus of attention is paramount to accurate probability judgment. Normatively, decision makers should judge the probability of a focal hypothesis by comparing it to all possible alternatives within its diagnostic category. However, rarely do people's probability judgments follow normative models. One factor contributing to the lack of normative responding is that only a small subset of items can be maintained in working memory at any point in time. Thus, the judged probability of a focal hypothesis is likely to be made by comparing it to only a subset of the alternatives, rather than the exhaustive set of alternative hypotheses.

There are several implications of our proposal that the number of hypotheses that one can maintain in the focus of attention is positively correlated with WM-capacity. One implication is that the judged likelihood of a focal hypothesis should decrease as the number and strength of relevant alternative hypotheses maintained in working memory increases, assuming that the focal is evaluated relative to its alternatives. Because WM-span and number of hypotheses generated are hypothesized to be positively correlated, judgments of the focal should also be negatively correlated with WM-span. We assume that participants partition the total probability of the sample space (cf. Fox & Rottenstreich, in press) across the subset of generated alternatives as if it were exhaustive, an assumption referred to as *constrained additivity* (this is because the partitioning ensures additivity among the generated alternatives). The partitioning of judged probability across the generated alternatives, rather than the exhaustive set, can lead to overestimation. For example, partitioning the total probability of 1.0 over three alternative hypotheses will lead each alternative to be overestimated if the exhaustive set includes eight alternatives. More generally, the more alternatives over which the total probability is partitioned, the lower the probability of each alternative (Fox & Rottenstreich, in press). Thus, to the extent to which participants fail to think of all alternative hypotheses, judgments of individual hypotheses should exceed the objective probabilities: Generating more alternatives should lead to lower individual probability judgments.

A second implication concerns the finding of subadditivity in probability judgment. Subadditivity is the tendency for the probability of an inclusive hypothesis to be rated as less than the sum of its components. For example, suppose one is asked to judge the probability that someone died from cancer given that they died from a natural cause (e.g., $p$(cancer | natural cause of death)), and later asked to judge the probability of death from various types of cancer given natural cause of death (e.g., $p$[breast cancer | natural death], $p$[leukemia | natural death], $p$[lung cancer | natural death], $p$[all other types | natural death]). Normatively, the probability assigned to the inclusive hypothesis is equal to the sum of its components. For example, $p$(cancer | natural cause of death) = $p$(breast cancer | natural death) + $p$(leukemia | natural death) + $p$(lung cancer | natural death) + $p$(all other types | natural death), because the constituents are mutually exclusive and exhaustive. However, considerable research has shown that the judged probability of the inclusive hypothesis is less than the sum of the judged probabilities of its components (for a review see Tversky & Koehler, 1994). That is, the inclusive hypothesis is *subadditive* with respect to the judged probabilities of the components (e.g., $p$[cancer | natural death] < $p$[breast cancer | natural death] + $p$[leukemia | natural death] + $p$[lung cancer | natural death] + $p$[all other types | natural death]).

We expected that the degree of subadditivity would be negatively correlated with WM-span because the judged probability of each individual hypothesis should decrease as a function of the number and strength of the alternative hypotheses generated. Note that this suggests that subadditivity should be negatively correlated with the overall strength of the generated alternatives—participants generating more, or stronger alternatives should be less subadditive than participants generating fewer, or less strong alternatives. In general, however WM-span should be negatively cor-

related with subadditivity because participants high in WM-span should be able to hold more alternatives in working memory.

We examined hypothesis generation and probability assessment using a simulated restaurant task. In the first phase of the task, participants viewed four regular restaurant customers each ordering one of eight items from one of four mutually exclusive menus. Each customer's menu selections over the course of 74 consecutive days were determined by a frequency distribution: Each participant saw distributions of 20–42–2–2–2–2–2–2 (breakfast), 20–20–20–3–3–3–3–2 (snack), 20–15–15–15–3–2–2–2 (dessert), and 20–10–9–9–8–8–8–2 (dinner), corresponding to the four menus. In the second phase of the experiment, participants judged the likelihood that each customer would order a randomly selected menu item on the 75th day. This judgment was followed by a thought-listing task (our measure of hypothesis generation) in which participants listed those items they considered when judging the focal item. Finally, at the end of the experiment, participants rated the likelihood of each of the 32 menu items one at a time. Subadditivity was calculated by summing each participant's probability judgments for the eight mutually exclusive and exhaustive items within each menu, which should sum up to 100% under the normative model. The four distributions constituted our independent variable and were predicted to affect the degree to which participants were subadditive (i.e., judgments exceeded 100%).

Assuming that participants generate only a few of the strongest alternatives, and that the judged probability of the focal is compared to the overall strength of the generated alternatives, then each individual probability judgment should be higher for the frequency distributions that are more evenly distributed (e.g., dessert and dinner) than for the frequency distributions where the majority of the frequencies are loaded on a few alternatives (e.g., breakfast and snack). This should lead to the greatest amount of subadditivity for the 20–10–9–9–8–8–8–2 distribution and the least amount for the 20–42–2–2–2–2–2–2 distribution. This prediction is easily seen by a simple calculation. Assume participants generate the two strongest alternatives for each focal judgment, and that each of the eight judgments within a distribution is made relative to the strength of the alternatives. Operationalizing strength as the objective frequency gives us 20/(20+42+2)+42/(20+42+2)+6(2/[20+42+2]) and a total sum of probabilities of 1.16 for the 20–42–2–2–2–2–2–2 distribution. For the 20–10–9–9–8–8–8–2 distribution, the corresponding sum of the probabilities is 20/(20+10+9)+10/(20+10+9)+2(9/[20+10+9])+3(8/[20+10+8])+2/(20+10+2) = 1.96. Hence, we predicted more subadditivity for the distributions that were more evenly distributed.

The above discussion gives rise to the following hypotheses. *H1a*: With respect to the effect of distribution, we hypothesized that subadditivity would increase as the distributions became more evenly distributed. Thus, the 20–10–9–9–8–8–8–2 distribution should yield the most subadditivity and the 20–42–2–2–2–2–2–2 distribution should yield the least. Based on WM limitations, we assume that participants will generate the same number of hypotheses for each distribution. If so, then the strength (sum of the objective frequencies) of the generated alternatives should also covary with distribution. This gives rise to *H1b*: We hypothesized that the sum of the objective frequencies of the generated alternatives would decrease as the distributions became more evenly distributed. This follows from our assumption that

participants generate the most likely alternatives. A necessary consequence is that the overall strength (i.e., objective frequency) of the generated alternatives will be the highest for the highest for the 20–42–2–2–2–2–2–2 distribution and the lowest for the 20–20–9–9–8–8–8–2 distribution. For example, the majority of the strength is captured by the two strongest alternatives (20, 42) in the first distribution, whereas the second distribution would require the generation of the four strongest alternatives to reach an approximately equivalent strength.

In addition to the effect of the distribution, we also hypothesized that both judgments of the focal and overall subadditivity would be related to individual differences in WM-capacity and to the number and strength of the generated alternatives: As WM-capacity increases, the number (and strength) of the alternatives used in the comparison process will increase, which, in turn should lead to lower probability judgment. That is, participants high in WM-capacity are hypothesized to generate more alternatives than participants low in WM-capacity. As a result, high-span participants should give lower probability judgments (and be less subadditive) than participants high in WM-capacity. Note that the overall strength of the alternative hypotheses (the sum of the objective frequencies) covaries with the number of alternatives (the more alternatives one thinks of the greater the overall strength). Thus, WM-capacity should also be related to the strength of the generated alternatives.

The above discussion gives rise to several hypotheses. First, judgments of the focal and overall subadditivity should be *H2a*: negatively correlated with the number of generated hypotheses, *H2b*: negatively correlated with the strength of the generated alternatives, and *H2c*: negatively correlated WM-span. Because the correlation between WM-span and judgments of the focal and overall subadditivity are assumed to arise because of the number and strength of the alternatives generated, we also hypothesized that: *H3a*: the number of alternatives generated would be positively correlated with WM-span, *H3b*: the strength of the generated alternatives would be positively correlated with WM-span, and *H3c*: the relationship between WM-span and subadditivity and between WM-span and single judgments should be mediated by the strength of the generated alternatives. That is, we expected that the relationship between WM-span and subadditivity (as well as focal judgments) would be mediated by the strength of the alternatives participants explicitly consider.

In addition to the above hypotheses, we can also test whether our assumption of constrained additivity is plausible. There are three aspects of the data that can be used to assess constrained additivity. First, the sum of the judgments of the generated alternatives and the focal should approximate 100%. This hypothesis is difficult to assess given the current experimental design, since it requires the assumption that participants always generate the same subset of hypotheses when judging each of the items listed in the thought-listing task. However, it is quite likely that participants will think of novel alternatives when rating the likelihood of the generated alternatives. Hence, failure to find precise constrained additivity cannot be taken as strong evidence against the assumption. However, there are two implications of our assumption of constrained additivity that should hold if it is indeed plausible. First, because we assume that participants partition the total probably over the subset of explicitly considered alternatives, the sum of the probabilities assigned to these al-

ternatives should not covary with WM-span. Although this is a null prediction, the deck is stacked against it. Recall that we also anticipate that participants high in WM-span will generate more alternatives. Thus, even though high-WM-span individuals should generate more alternatives, the sum of the judgments assigned to the generated alternatives should be uncorrelated with WM-capacity. The second implication of constrained additivity is that the sum of the probabilities assigned to the explicitly considered alternatives should be unaffected by distribution. Thus, we predicted a null effect of distribution on constrained additivity, despite the fact that we expected overall subadditivity to be affected by distribution.

## 2. Method

### 2.1. Participants

Participants were 40 undergraduates at the University of Maryland enrolled in Psychology courses. For participating they received partial credit in their course.

### 2.2. Procedure

Prior to beginning the experiment, each participant completed the operation-span task (Conway & Engle, 1996; Turner & Engle, 1989), as a measure of WM-span. This task requires participants to retain a list of words while solving mathematical problems. For example, on successive presentations participants might see (6*2)−3 = 4 ? DOG; (8/2)−3 = 1 ? WINDOW, etc. Participants read aloud the equation, responded yes or no to whether the equation was correct, and then read the word aloud. After stating the word, the experimenter proceeded to the next operation–word pair. This continued until the participant was prompted to recall the words in the order in which they were presented. Participants were presented with 15 sets of equation-words pairs, with each set size (2, 3, 4, 5, 6) presented three times in random order. Participants' WM-span scores were calculated by summing up the number of words recalled in the correct serial position (for a description of the operation-span task see Turner & Engle, 1989). Thus, the maximum possible score was 60 if participants correctly recalled all words from the 15 lists perfectly. Klein and Fiss (1999) reported that the test–retest reliability of the O-span task ranged from $r = 0.667$ (2–3 week intertest interval) to $r = 0.812$ (6–7 week intertest interval). Our split-half reliability approached their lower bound, $r = 0.59$, $p = 0.0002$, with a Cronbach's alpha coefficient of 0.74, indicating reasonably good reliability.

The experimental task provided a simulation of ''days'' in a restaurant. There were 32 total food items with eight items on each of four different menus. These menus were breakfast (pancakes, eggs, etc.), snack (almonds, popcorn, etc.), dinner (steak, chili, etc.), and dessert (cake, ice cream, etc.). Four ''regular'' customers were seen each day, each always ordering from one menu (e.g., only Bob ordered from the breakfast menu). Each day each customer ordered one of the eight menu items from their respective menus. Their choice of menu items was determined by a frequency

distribution in which the items were presented at differing relative frequencies. Over the four menus, each participant saw distributions of 20–42–2–2–2–2–2–2 (breakfast), 20–20–20–3–3–3–3–2 (snack), 20–15–15–15–3–2–2–2 (dessert), and 20–10–9–9–8–8–8–2 (dinner).

Participants proceeded through a series of "days" during which they saw each of the regular customers and their orders. Each day proceeded in the following order: (1) breakfast displayed with a cartoon image of Bob, an item from the breakfast menu and the word "Breakfast," (2) Steve, a snack item and "Snack", (3) Tim, a dinner item and "Dinner", (4) Dan, a dessert item and "Dessert," (5) a picture of a sunset signifying the end of the day. Participants were presented with 296 menu items over 74 simulated days. Each menu item was presented for 3 s, with the order of the items within each menu random for each participant. To ensure that participants were attending to the task, eight times during the learning phase they were prompted to recall the most recent item ordered by a particular customer.

After viewing 74 simulated days, participants judged the probability of each customer ordering a particular menu item (e.g., Given what you have seen so far, what is the probability that Bob will have pancakes on the next day?). One menu item was chosen randomly from each distribution for each participant (i.e., participants rated one item per distribution). In fact, this initial judgment task was done to assess hypothesis generation—which, and how many, alternative hypotheses participants thought of when prompted to make a probability judgment. Thus, after entering each judgment, participants engaged in a thought-listing task (Cacioppo & Petty, 1981) in which they typed into the computer the items they considered while making the probability judgment. [1] Participants then judged the probability of each of the listed items in isolation of the others. This task was repeated four times, once for each distribution (menu), with order of distribution randomized. At the end of the experiment, participants rated the probability of each of the 32 menu items one at a time in random order without engaging in the thought-listing task. These 32 judgments (eight per meal) enabled us to assess the degree to which participants' judgments were subadditive. The subadditivity score was calculated by summing up the judgments for each set of eight items, which should normatively sum to 100. The scale used for all judgments was an 11-point scale ranging from 0%—impossible to 100%—certain. Care was taken to point out that the judgment should be made by considering how often the item had occurred throughout the entire experiment, not on what was ordered most recently. This instruction was emphasized during a practice session, as well as prior to the first judgment phase in the restaurant task.

Prior to engaging in the experimental task, participants completed a practice session that resembled the experimental task in all important aspects. The practice session was much shorter than the restaurant task and involved learning, then

[1] Because participants performed the thought-listing task four times throughout the course of the experiment, the reliability of our measure of hypothesis generation can be assessed by examining the correlations between number of alternatives generated for the four different distributions. These correlations ranged from $r = 0.69$ to $r = 0.84$, all of which were statistically significant at the $p < 0.0001$ level. The overall Cronbach alpha coefficient was 0.93.

estimating the likelihood of, where two travelers went for vacation. The practice task was used to introduce participants to the probability judgment and thought-listing components of the experimental task.

## 3. Results

Four participants were excluded from the analyses for failing to perform the thought-listing task.

### 3.1. Subadditivity, hypothesis generation, and the effect of distribution

We hypothesized that subadditivity would be the highest for distributions that contained primarily weak (low frequency) alternatives and lowest for distributions that contained several strong (high frequency) alternatives (*H1a*). Thus, if participants can hold the focal plus two alternatives, and they generate the strongest alternatives, each individual focal judgment should be lower for the distributions with stronger alternatives, resulting in less overall subadditivity for those distributions. In contrast, if participants maintained all eight items in the focus of attention, judgments should approximate additivity (sum up to 100%) and there should be no effect of distribution.

Table 1 shows the mean number of alternatives generated, the mean sum of the judgments, the overall *strength* of the generated alternatives (i.e., objective relative frequencies of the generated alternatives), and the constrained additivity scores. The number of alternatives generated did not differ across distributions: Participants generated roughly three alternatives regardless of distribution, $F(3, 96) = 0.28$, $p > 0.05$. In contrast, the subadditivity score differed considerably across the distributions, and in the predicted direction, with the greatest amount of subadditivity for the distributions that had primarily weak (low frequency) alternatives, $F(3, 105) = 9.66$,

Table 1
Means (standard errors) for major dependent variables for the four distributions

| | Breakfast 20–42–2–2–2–2–2–2 | Snack 20–20–20–3–3–3–3–2 | Dessert 20–15–15–15–3–2–2–2 | Dinner 20–10–9–9–8–8–8–2 |
|---|---|---|---|---|
| Number of alternatives generated | 3.0 (0.3) | 2.9 (0.3) | 3.3 (0.3) | 3.1 (0.3) |
| Strength of generated alternatives | 50.3 (3.3) | 44.0 (2.5) | 40.4 (3.3) | 37.0 (3.0) |
| Subadditivity score | 238 (18.9) | 245 (16.2) | 273 (21.9) | 294 (21.7) |
| Constrained additivity | 125 (8.9) | 133 (12.6) | 129 (8.9) | 126 (9.7) |

*Note*: Number of alternatives generated = number of alternative menu items listed in the thought-listing task. Strength of the generated alternatives = sum of the objective frequencies of the items generated in the thought-listing task. Subadditivity score = sum of the probability judgments given for all eight items. Constrained additivity = sum of the probability judgments assigned to the focal and alternatives generated in the thought-listing task.

$p < 0.001$. A priori pairwise comparisons using Holm's adjustment revealed that only the two most dissimilar distributions (breakfast and dinner) differed statistically from each other; however the ordering of all four distributions follows the shape of the frequency distributions: Distributions with stronger (higher frequency) alternatives elicited less subadditivity. Note that nearly 85% of the objective frequencies are loaded on the two strongest alternatives (20, 42) in the breakfast distribution, but that only 40% of the objective frequencies are loaded on the two strongest alternatives (20, 10) in the dinner distribution. Thus, if judgments are based on a comparison between the focal and the few strongest alternatives, judgments should be lower in the breakfast distribution than in the dinner distribution. This would lead to less subadditivity for the breakfast distribution, as was found. The remaining two distributions, snack (20–20–20–3–3–3–3–2) and dessert (20–15–15–15–3–2–2–2), showed intermediate levels of subadditivity, with less subadditivity in the snack distribution, as expected.

Embedded in our explanation of why subadditivity decreased as a function of the strength of the alternatives is the assumption that participants generated the few strongest alternatives (those which had the highest objective frequency). To examine this directly, we classified alternatives with an objective frequency exceeding 9.25 (the mean objective frequency) as high frequency and those below the mean as low frequency and tallied the proportion of high-frequency and low-frequency items participants generated. For example, the Breakfast distribution had two high-frequency alternatives (20 and 42 are greater than 9.25), while the Dessert distribution had four high-frequency alternatives. Thus, for each participant the proportion of the possible high- and low-frequency items that were generated was calculated (i.e., generating both high-frequency items in the Breakfast category would yield a proportion of 1.0). Collapsing across distribution, the mean proportion of the high-frequency items generated ($M = 0.53$) exceeded the mean proportion of the low-frequency items generated ($M = 0.31$), $t(35) = 6.43$, $p < 0.001$. Thus, in general, participants tended to generate strong alternative hypotheses.

We also calculated the strength (i.e., objective relative frequencies) of the generated alternatives for each distribution. If the finding of subadditivity is related to the overall strength of the generated alternatives, then the overall strength of the generated alternatives should be highest in the distribution with the lowest subadditivity (*H1b*). This was indeed the case, $F(3, 87) = 6.26$, $p = 0.001$. Whereas subadditivity increased as the number of strong alternatives decreased, the overall strength of the generated alternatives decreased (see Table 1). Note that there were no differences in the *number* of alternatives generated across the distributions (roughly three alternatives for all four distributions), suggesting that participants are considering the strength of the generated alternatives, not merely the number of alternatives, when forming their judgments. This is consistent with prior research by Dougherty et al. (1997) who showed that the judged probability of a focal was higher when the alternatives were improbable.

The above results indicated two main findings: (1) Participants tended to generate the strongest few alternatives, and (2) the strength of the generated alternatives affected the degree to which participants were subadditive. We now examine the role of individual differences in WM-span in hypothesis generation and subadditivity.

### 3.2. WM-span, hypotheses generation and probability judgments

Our main hypotheses focused on the correlations among WM-span, number of alternatives generated, the strength of the generated alternatives, and three components of the probability estimation task: (a) single-item judgments, where the single-item judgments were the four items (one per distribution) for which participants engaged in the thought-listing task, (b) overall subadditivity (the sum of the probabilities assigned to the eight menu items within each distribution), and (c) the constrained additivity score (the sum of the probabilities assigned to the generated alternatives for each of the four distributions). These correlational analyses were done collapsing across the four distributions.

We predicted that the strength and/or number of alternatives generated would be negatively correlated with both single judgments and overall subadditivity (*H2a*, *H2b*). That is, generating several strong alternatives should lead to lower judged probability and less subadditivity than generating several weak alternatives. We also predicted that WM-capacity would be positively correlated with the number and strength of the alternative hypotheses generated (*H3a*, *H3b*). Finally, we hypothesized that WM-span would be related to single judgments and subadditivity, but that this correlation would be mediated by the strength of the generated alternatives (*H3c*).

Table 2 presents the correlations among the main variables. As can be seen, both predictions were supported, with all the significant correlations corresponding to medium effects or larger (Cohen, 1988). [2] The correlation between WM-capacity and the average number of hypotheses generated was positive and significant, $r = 0.45$, $p = 0.006$. Comparison of the upper 25 percentile (high span) with the lowest 25 percentile (low span) revealed that high-span participants generated an average of 4.1 alternatives whereas low-span participants generated only 2.4 alternatives. By Cohen's $d$, the effect size was quite large, $d = 1.0$, indicating that a full standard deviation separated the high- and low-span participants. Thus, participants high in WM-span generated more alternatives. In addition, the overall strength of the generated alternatives was negatively correlated with both the amount of overall subadditivity, $r = -0.37$, $p = 0.03$, and the judgments of the single hypotheses, $r = -0.41$, $p = 0.01$. However, the correlations between number of alternatives generated (ignoring strength) and judgments of single hypotheses and subadditivity were both negative, but non-significant. Note that the overall strength of the generated alternatives was highly correlated with number generated ($r = 0.92$), but that only the former was significantly correlated with subadditivity and judgments of single hypotheses. This makes sense since participants should take into account the strength of the alternatives, not just how many they generated: Generating three highly likely alternatives should reduce the judged likelihood of the focal more than generated three unlikely alternatives.

[2] According to Cohen (1988), $r = 0.30$ is considered a medium effect size.

Table 2
Pearson $r$ correlations between major variables

| | WM-span | Single judgment | Number of alternatives | Judgments of generated alternatives | Overall subadditivity |
|---|---|---|---|---|---|
| Single judgment | −0.25 | | | | |
| Number of alternatives | 0.45** | −0.30 | | | |
| Judgments of generated alternatives | 0.07 | −0.37* | 0.57** | | |
| Overall subadditivity | −0.37* | 0.84** | −0.25 | 0.28 | |
| Objective sum of generated alternatives | 0.42* | −0.41* | 0.92** | 0.50** | −0.37* |

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

We also hypothesized that the amount of subadditivity would be inversely related to WM-capacity (*H2c*), since WM-capacity constrains the number of alternatives one can maintain in the focus of attention. This prediction was supported: Overall subadditivity correlated negatively with WM-span, $r = -0.37$, $p = 0.02$. A comparison of high-span (upper 25%) and low-span (lowest 25%) participants revealed that the sum of probability judgments for high-span participants ($M = 222.5$) was over 100% lower than the sum of the probability judgments for low-span individuals ($M = 335.2$), indicating that high-span participants were much less subadditive. Again, by Cohen's $d$ this corresponded to a relatively large effect size, $d = 0.81$.

Using mediation analysis (Baron & Kenny, 1986), we tested whether the relationship between WM-span and judgments of the single hypotheses (and subadditivity) was due to the strength of the generated alternatives (*H3c*). There are three steps to mediation analyses. First, there must be a significant relationship between the predictor or independent variable (WM-span) and the dependent variable (judged probability or subadditivity), though this step is not critical for establishing mediation. Second, there must be a significant relationship between the predictor variable (WM-capacity) and the potential mediator (strength of generated alternatives). Finally, the effect of the mediator (strength of the generated alternatives) on the dependent variable (judged probability or subadditivity) should be significant when controlling for the predictor variable (WM-capacity). Full mediation occurs when the predictor is no longer a significant predictor of the dependent variable when controlling for the mediator.

The first mediation test examined whether the strength of the generated alternatives mediated the relationship between WM-capacity and judgments of the focal events (*H3c*). This set of analyses revealed that the direct relationship between WM-capacity (the predictor variable) and judged probability (the dependent variable) was non-significant, $b = -0.45$, $t = 1.50$, $p = 0.14$. Although this first step was non-significant, Baron and Kenny (1986) point out that mediation is still possible in such cases. Thus, proceeding with caution, we examined the effect of WM-span on the strength of the generated alternatives and the effect of the generated alterna-

tives when controlling for WM-span. This set of analyses revealed a significant relationship between WM-capacity and the strength of the generated alternatives ($b = 0.71$, $t = 2.69$, $p = 0.01$), and a marginally significant effect of strength of the generated alternatives on focal judgments when controlling for WM-span ($b = -0.31$, $t = 1.65$, $p = 0.10$). Using a Sobel test, the mediated path was found to be in the predicted direction, but not statistically significant, $z = -1.41$, $p = 0.157$.

A similar mediation analysis was performed to test the hypothesis that the strength of the alternatives generated mediated the relationship between WM-capacity and subadditivity (*H3c*). In this case, there was a significant relationship between WM-capacity and overall subadditivity ($b = -4.58$, $t = 2.35$, $p = 0.02$), and between WM-capacity and strength of the alternatives generated ($b = 0.71$, $t = 2.69$, $p = 0.01$). However, the relationship between strength of the generated alternatives and subadditivity when controlling for WM-capacity failed to reach conventional significance ($b = -1.89$, $t = 1.51$, $p = 0.14$). The test of mediation was non-significant using Sobel's approach, $z = -1.31$, $p = 0.189$, though the result was in the predicted direction. [3] It is worth noting that WM-capacity also was not a significant predictor of subadditivity when strength of the generated alternatives was controlled, $b = -3.25$, $t = 1.54$, $p = 0.13$.

In light of the failure to find conclusive evidence of mediation, one might ask whether the effect of distribution is due to the strength of the generated alternatives. That is, suppose we were to equate each distribution on the strength of the alternatives generated. Would this lead to a decrease in the effect of distribution on subadditivity? We showed above that the different distributions affected both subadditivity and the strength of the generated alternatives, but in the opposite directions: The lowest amount of subadditivity was observed for the distributions in which the strength of the generated alternatives was greatest. This suggests that the effect of distribution is due to the strength of the generated alternatives within each distribution. To test this more directly, one can examine whether the effect of distribution is statistically significant after the distributions have been equated on the strength of the generated alternatives.

To examine whether the effect of distribution on subadditivity is reduced after equating the distributions on the strength of the generated alternatives, we ran two statistical tests. One examining the effect of distribution by itself, and one

[3] Although the above tests failed to reveal mediation, it is important to point out that our test likely lacked statistical power. As Baron and Kenny (1986) point out, the lack of statistical power is of particular concern when the path between the predictor (WM-span) and the mediator (strength of the generated alternatives) is stronger than the path between the mediator (strength of the generated alternatives) and the dependent variable (judged probability or subadditivity). This problem, called multicollinearity, compromises the power of the statistical test because there is little unique variance to be explained by the mediator (strength of the generated alternatives) once the variance due to the predictor (WM-capacity) is removed. In fact, one actually has more statistical power when the relationship between the predictor and the mediator is relatively weak. As one can readily see in Table 2, the relationship between WM-span and the strength of the generated alternatives ($r = 0.42$) was actually greater than the relationship between the strength of the generated alternatives and judged probability ($r = -0.41$) and subadditivity ($r = -0.37$).

examining the effect of distribution using the strength of the generated alternatives as a covariate. Whereas the effect of distribution was significant by itself, $F(3,105) = 9.66$, $p = 0.0003$, its effect was much reduced when the strength of the alternatives was included as a covariate, $F(3,102) = 3.31$, $p = 0.02$. Examination of the effect size of distribution yielded $R^2 = 0.275$ when the strength of the alternatives was not controlled for, but $R^2 = 0.097$ when variance due to the strength of the alternatives was removed. One way to conceptualize this comparison is that equating the four distributions on the strength of the generated alternatives reduces the effect of distribution on subadditivity. This suggests that the strength of the generated alternatives mediated the effect of distribution on subadditivity.

### 3.3. Constrained additivity

We assumed that participants partitioned the total probability over the subset of hypotheses contained in working memory (the focal plus the alternatives generated)—the assumption of constrained additivity. There are three implications of the constrained additivity assumption. First, the sum of the judgments assigned to explicitly considered alternatives (the focal plus alternatives generated) should approximate 100%. However, in order to test this we would have to assume that participants considered the same subset of alternatives when judging each item listed in the thought-listing task. This is unlikely to be the case, since participants will undoubtedly think of new alternatives throughout the judgment task. Nevertheless, examination of the mean constrained additivity scores (see Table 1) showed that they approximated 100%, showing slight, albeit significant overestimation.

The second implication of constrained additivity is that the sum of the judgments for the generated alternatives should not be affected by distribution. That is, if participants partition the probabilities over the set of alternatives considered, the sum of the alternatives over which the partitioning takes place should not depend on the distribution of the alternatives. Indeed, as expected there was no effect of distribution on the sum of these probabilities, $F(3,30) = 0.25$, $p = 0.86$ (see Table 1 for means).

Finally, the third implication of constrained additivity is that there should be no relationship between the sum of the judgments of the generated alternatives and WM-span. Note that the deck is stacked against this prediction since we expected, and found, that WM-span was positively correlated with number of alternatives generated. Thus, even though high-WM-span participants generated more alternatives, we did not expect the sum of probabilities assigned to the generated alternatives to covary with WM-span. The correlation between the additivity of the generated alternatives and WM-span was non-significant, $r = 0.07$, $p = 0.70$, indicating that while WM-capacity influenced the number of alternatives generated, the sum of the judgments for the generated alternatives was independent of capacity. Although we cannot determine whether strict constrained additivity holds, the above two results lend support for the assumption.

## 4. General discussion

Our research has directly implicated working memory as an important process in hypothesis generation and probability judgment. Indeed, the present experiment revealed three main findings: (1) the judged probability of a single hypothesis, and the degree to which participants were subadditive, were both negatively related to the strength of the generated alternatives and WM-span, (2) participants tended to generate the strongest alternatives, and (3) the number and strength of the alternatives generated was constrained by WM-capacity. These findings are important for two reasons: First, our findings indicate that individual differences in WM-capacity are fundamental to hypothesis generation and probability assessment. Second, and perhaps more important, our research points to a greater need for exploring issues of capacity limitations in the context of judgment and decision making tasks. Although capacity limitations have long been assumed to affect judgment, relatively few studies have explored the issue from within the context of working memory theory.

One possible criticism of the present research is that our use of the thought-listing procedure might have prompted participants to be more reflective and possibly generate alternatives when they might not naturally be inclined to generate alternatives. That is, perhaps the thought-listing task enticed participants to use a more consciously controlled search of memory for alternatives than would be used in the absence of the thought-listing procedure. Thus, the inclusion of the thought-listing procedure might be responsible for driving the correlations among subadditivity, hypothesis generation, and WM-span. Although this criticism may hold, it does not undermine the validity of our results. Even if the thought-listing procedure did induce a more controlled search of memory, it would merely place boundary conditions on when we would expect to find a relationship between WM-span and subadditivity: The relationship would be restricted to those tasks in which people use controlled memory search processes.

This said, we doubt that the thought-listing procedure, as implemented in our experiment, affected how people judged probability in the probability estimation phase of our experiment. Participants engaged in the thought-listing task in a separate phase of the experiment than the judgment task on which the subadditivity scores were based. There was absolutely no encouragement for participants to think of alternatives while they were judging the exhaustive set of 32 menu items.

One obvious problem with making inferences based on correlational data is the third variable problem. We proposed that the relationship between WM-span and probability judgment, in particular subadditivity, was due to the number and strength of the alternatives used in the comparison process. According to our hypothesis, participants should be less subadditive as the number of alternatives used in the comparison process increases. We proposed that WM is the underlying cognitive construct that governs how many alternatives participants can maintain. However, there is one other possibility that might account for our results—the results may reflect crystallized intelligence. It is possible that participants high in WM-span have knowledge that the sum of the probabilities of a set of mutually exclusive and exhaustive hypotheses should be additive. Thus, it is possible that the correlation

between WM-span and subadditivity is the crystallized intelligence about probability theory. We refer to this possibility as the *knowledge hypothesis*.

The combined outcome of the correlation analyses and the effect of distribution argue against the knowledge hypothesis as the sole factor underlying our findings. One test that sheds light on the validity of the knowledge hypothesis is the test of the interaction between WM-span and distribution. There are two possible reasons that WM-span might interact with distribution. One reason is that the incremental reduction in subadditivity should be greater for the even (20–10–9–9–8–8–8–2) distribution than for the uneven (20–42–2–2–2–2–2–2) distribution. That is, considering an extra alternative that had occurred with frequency 2 in the learning task will lead to less of a reduction in probability judgment than considering an alternative that occurred with frequency 8. Because we assume that high-span participants include more alternatives in the comparison process, we would also expect there to be a greater effect of WM-span on subadditivity as the distribution of alternatives became more evenly distributed. Note, however, for this account to hold, we have to assume that participants retrieve a constant number of alternatives across the four distributions. Note also that the relative accessibility of the alternatives in the four distributions is confounded with the objective frequency. Thus, it is possible that participants might retrieve more alternatives in the 20–10–9–9–8–8–8–2 distribution simply because the alternatives are more accessible. Although our data showed that participants retrieved the same number of alternatives in the thought-listing task, we do not know if this result generalized to the main probability judgment task. If the relative accessibility affected the number of hypotheses recalled, it would mitigate the interaction effect.

The second possible reason for a WM-span $\times$ distribution interaction is related to the knowledge hypothesis. That is, if the reason WM-span is correlated with subadditivity is that high-span participants have knowledge that their judgments should sum to 100, then their judgments should be unaffected by the distribution of the alternatives because participants should use their knowledge of additivity regardless of the distribution of the alternatives. Thus, if high-span participants have knowledge that judgments should be additive, one would expect additivity (or approximate additivity) to hold for all four distributions. Analyses of subadditivity across the four distributions using WM-span as a quantitative predictor failed to reveal a WM-span by distribution interaction ($p = 0.62$). Whereas the lack of the interaction does not lend support for the theory that subadditivity is due to limitations in WM-capacity, it provides counterevidence to the hypothesis that the correlation between WM-capacity and judgment was due to knowledge of probability theory.

Particularly diagnostic for dispelling the knowledge hypothesis was the finding that only two participants in our study produced judgments that were additive. One of these participants was additive for only one of the four distribution and had a mean sum of 115. The other participant was additive for three of the four distributions and had a mean sum of 102.5. The participant who was additive for one distribution had a WM-span score of 28, and the participant who was additive for three distributions had a WM-span score of 16 (which was the fourth lowest score in the distribution). Note that the mean and median WM-span scores were $M = 25.00$

(STD = 8.89) and $X_{50} = 25.5$, respectively, and that the distribution ranged from 7 to 48. Neither of the two participants who showed evidence of additivity was in the upper quartile of the WM-span distribution. Although this argues against knowledge of probability theory as an interpretation of our results, it is still possible that some participants knew that their judgments should be additive and attempted to make them so. However, our data would imply that knowledge of additivity did not covary with WM-span in our experiment, or that knowledge of probability theory was not sufficient to circumvent the effect of WM-limitations on probability judgment.

One might question whether the observed relationship between WM-span and subadditivity reflects retrieval processes rather than a limitation in the number of alternatives that can be maintained in the focus of attention. That is, perhaps it is the case that participants high in WM-span simply are more efficient at generating or retrieving alternatives from long-term memory. We do not view this possibility as an alternative explanation of our findings, but rather as part of the overall explanation. Indeed, prior research by Rosen and Engle (1997) using a verbal-fluency task revealed that high-span participants were better able to inhibit the generation of irrelevant category exemplars and consequently showed higher fluency scores than low-span participants. Rosen and Engle argued that the enhanced fluency scores for high-span participants was due to the their ability to devote some (untapped) WM-resources to inhibiting the resampling (or regeneration) of already generated items. In contrast, low-span participants did not have unused WM-resources to devote to inhibition. In support of this hypothesis, Rosen and Engle showed that high-span participants, but not low-span participants, performed significantly worse when the fluency task was done under divided attention conditions.

In the context of our experiment, it is possible that the difference between high-span and low-span participants was that the high-span participants were better able to inhibit already generated, or irrelevant alternatives (i.e., those from a non-target distribution) from taking-up space in the WM store. In either case, however, the net result would be that WM-span determined the number of *relevant* alternatives participants considered. Although number of relevant alternatives used in the comparison process and the ability to inhibit irrelevant alternatives from replacing relevant alternatives in WM probably reflect the same underlying process, it might be possible to separate these two possibilities experimentally. As mentioned above, the work by Rosen and Engle (1997) suggests that high-span participants are effective at inhibiting the regeneration of exemplars in tests of fluency. However, this inhibition ability is mitigated in high-span participants when given a secondary task to perform. In contrast, low-span participants are unaffected by secondary task loads. If inhibition plays a role in hypothesis generation tasks, it may be possible to show that hypothesis generation and judgment in high span, but not low span, participants are affected by manipulations of concurrent WM loads. We are currently testing this possibility.

### *4.1. Implications for prior research on probability judgment*

Our results have implications for several lines of research on probability judgment. Windschitl and Wells (1998; Windschitl & Young, 2001) illustrated that the

judged probability of the focal is often made by comparing its strength to the strongest single alternative—a heuristic they referred to as the comparison heuristic. This contrasts with the normative view, which requires that the focal be judged by comparing it to all possible alternatives. Our conceptualization of the comparison process places it somewhere between the normative view and the view of Windschitl and Wells: Rather than considering all possible alternatives, or considering merely the single strongest alternative, our data suggest that the number of alternatives to which the focal is compared is greater than one, but has an upper limit that depends on one's WM-capacity.

A second line of research for which our results have implications is research done from within the framework of support theory (Tversky & Koehler, 1994). Support theory proposes that people compare focal hypotheses with an aggregate (or packed) set of the alternatives. Subadditivity is assumed to result from underweighting (or underestimating) the evidential support for the packed set of hypotheses (Tversky & Koehler, 1994). In essence, people were assumed to overestimate probabilities because they failed to generate adequate support for alternative hypotheses. Our research provides a mechanism by which support might be generated. Rather than treating the packed hypothesis as a composite, and underweighting the support for the composite, our results suggest that participants partially unpack the set of alternative hypotheses by generating the most likely alternatives, and that this generation, or unpacking, is constrained by one's WM-capacity. By this account, subadditivity results, at least in part, from the failure to completely unpack, or consider, the set of alternative hypotheses.

### *4.2. Summary*

In sum, the idea of a limited capacity working memory has had a lasting impact on a number of areas of cognitive psychology, including judgment and decision making. However, despite its obvious importance in judgment and decision making tasks, judgment and decision researchers have been slow to incorporate models of working memory into models of judgment. Instead, decision researchers "have mostly made use of memory limitations to motivate a concern for issues of strategy selection" (Weber, Goldstein, & Barlas, 1995, p. 35). Indeed, relatively little research, in fact no published report that we could find, has directly examined the role of working memory, or capacity limitations, in probability judgment or hypothesis generation. The present research, and prior research in our lab (Dougherty, 2001; Dougherty, Gettys, & Ogden, 1999), indicates the need for greater application of memory theory in the area of judgment and decision making.

## Acknowledgements

This research was supported by funds provided by the University of Maryland Department of Psychology, the University of Maryland Graduate Research Board, and Grant SES-0134678 from the National Science Foundation.

## References


Barrows, H. S., Norman, G. R., Neufeld, V. R., & Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general medicine practice. *Clinical and Investigative Medicine, 5*, 49–55.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 35*, 29–46.

Cacioppo, J. T., & Petty, R. T. (1981). Social psychological procedures for cognitive response assessment: the thought-listing technique. In T. V. Merluzzi, C. R. Glass, & M. Benest (Eds.), *Cognitive assessment* (pp. 309–342). New York: Guilford.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Conway, A. R. A., & Engle, R. W. (1996). Individual differences in working memory capacity: more evidence for a general capacity theory. *Memory, 4*, 577–590.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–185.

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

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). Minerva-DM: a memory processes model for judgments of likelihood. *Psychological Review, 106*, 180–209.

Dougherty, M. R. P., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes, 70*, 135–148.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: an analysis of clinical reasoning*. Cambridge: Harvard U P.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake, & P. Shah (Eds.), *Models of working memory: mechanisms of active maintenance and executive control* (pp. 102–134). New York: Cambridge.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General, 128*, 309–331.

Fisher, S. D., Gettys, C. F., Manning, C., Mehle, T., & Baca, S. (1983). Consistency checking in hypothesis generation. *Organizational Behavior and Human Performance, 31*, 233–254.

Fox, C. R., & Rottenstreich, Y. Partition priming in judgment under uncertainty. *Psychological Science*, in press.

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Decision Processes, 24*, 93–110.

Gettys, C. F., Pliske, R. M., Manning, C., & Casey, J. T. (1987). An evaluation of human act generation performance. *Organizational Behavior and Human Decision Processes, 39*, 23–51.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging. A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation: advances in research and theory: Vol. 22* (pp. 193–225). San Diego: Academic Press.

Kane, M. J., Bleckley, M. K., Conway, A. R., & Engle, R. A. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130*, 169–183.

Klein, K., & Fiss, W. H. (1999). The reliability and stability of the Turner and Engle working memory task. *Behavioral Research, Methods, Instruments, and Computers, 31*, 429–432.

Koehler, D. K. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110*, 449–519.

Koehler, D. K. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology, 20*, 461–469.

Libby, R. (1985). Availability and the generation of hypotheses in analytical review. *Journal of Accounting Research, 23*, 648–667.

Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica, 52*, 87–106.

Mehle, T., Gettys, S., Manning, C., Baca, S., & Fisher, S. (1981). The availability explanation of excessive plausibility assessments. *Acta Psychologica, 49*, 127–140.

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General, 126*, 211–227.

Sanbonmatsu, D. M., Posavac, S. S., Kardes, F. R., & Mantel, S. P. (1998). Selective hypothesis testing. *Psychonomic Bulletin and Review, 5*, 197–220.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*, 127–154.

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

Weber, E. U., Böckenholt, U., Hilton, D. J., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1151–1164.

Weber, E. U., Goldstein, W. M., & Barlas, S. (1995). And let us not forget memory: the role of memory processes and techniques in the study of judgment and choice. In J. Busemeyer, R. Hastie, & D. L. Medin's (Eds.), *Decision making. The psychology of learning and motivation: advances in research and theory: Vol. 32* (pp. 33–81). San Diego, CA: Academic Press.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology, 75*, 1423–1441.

Windschitl, P. D., & Young, M. E. (2001). The influence of alternative outcomes on gut-level perceptions of certainty. *Organizational Behavior and Human Decision Processes, 85*, 109–134.

Windschitl, P. D., Young, M. E., & Jensen, M. E. (2002). Likelihood judgment based on previously observed outcomes: the alternative-outcomes effect in a learning paradigm. *Memory and Cognition, 30*, 469–477.