Memory & Cognition
2003, 31 (6), 968-982

Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval

MICHAEL R. P. DOUGHERTY and JENNIFER HUNTER
University of Maryland, College Park, Maryland

In this research, we examined the role that individual differences in working memory (WM) capacity, the strength of alternatives, and time constraints play in probability judgment and subadditivity. With a laboratory-based learning task, Experiment 1 revealed that the degree to which participants' probability judgments were subadditive was negatively correlated with a measure of WM capacity, even when variance due to short-term memory capacity was removed. In addition, participants were more subadditive when the viable alternatives were all rather weak. Experiment 2 extended the WM-capacity–subadditivity correlation to a population judgment task and revealed that subadditivity increases when the judgment task is performed under time constraints. Results support a model that assumes that people make probability judgments by comparing the focal hypothesis with relevant alternatives retrieved from long-term memory and that people high in WM span include more alternatives in the comparison process. Time constraints are assumed to truncate the alternative generation process, leading to fewer alternatives being recalled from long-term memory.

A fundamental assumption of many models of human probability judgment is that judgments are made by comparing the focal hypothesis with one or more alternative hypotheses. For example, Windschitl and Wells (1998) proposed that participants judge the focal hypothesis by using a comparison heuristic in which the strength of the focal hypothesis is compared with the strongest single alternative. Tversky and Koehler's (1994) *support theory* proposed that people judge the focal hypothesis by assessing the balance of evidential support in favor of the focal hypothesis versus alternative hypotheses. Finally, Dougherty, Gettys, and Ogden's (1999; Dougherty, 2001) Minerva-DM (MDM) model assumed that the probability of an individual hypothesis is given by the ratio of the memory strength of the focal hypothesis to the memory strengths of the alternative hypotheses. Although these theories propose slightly different mechanisms for how probability judgments are made, all explicitly assume that the to-be-judged hypothesis is compared with at least one alternative hypothesis.

The nature of the comparison process has important consequences for the accuracy of probability judgments. For example, Windschitl and Wells (1998) found that verbal probability judgments were sensitive to the distribution of alternative hypotheses. In one study, participants played a hypothetical lottery with five other "players" and were asked to estimate the likelihood that they would win the lottery, given that they held 21 tickets. In one condition, the five other players held 14, 13, 15, 12, and 13 tickets. In another condition, the five players held 52, 6, 2, 2, and 5 tickets. The participants' judgments were lower for the second distribution than for the first even though the objective probability of the participant's winning was .24 for both conditions (e.g., they held 21 out of 88 total tickets). Windschitl and Wells argued that judgments were lower for the second distribution because the participants compared the number of tickets they held with the number of tickets held by the dominant alternative. Thus, according to Windschitl and Wells, the judged probability of the focal hypothesis depends on a comparison with the strength of the strongest alternative (see also Windschitl & Young, 2001; Windschitl, Young, & Jensen, 2002).

Windschitl and Wells's (1998) comparison heuristic is a special case of Tversky and Koehler's (1994) support theory. According to support theory, judged probability is given by the ratio of evidential support for the focal hypothesis versus the alternative hypothesis:

$$p(\mathrm{A,B}) = \frac{s(\mathrm{A})}{s(\mathrm{A}) + s(\mathrm{B})}, \tag{1}$$

where p(A, B) is the probability of Hypothesis A rather than B and s(A) and s(B) are the support in favor of A and B, respectively. Consider the judgment p(rain to-

This material is based on work supported by the National Science Foundation under Grant SES-0134678 and by the University of Maryland Graduate Research Board. The authors thank Brandon Abbs for assisting in data collection and Rickey Thomas, Thomas Wallsten, Kent Norman, Ana Franco-Watkins, and Amber Lehman for providing feedback on earlier drafts of this article. Correspondence concerning this article should be addressed to M. R. P. Dougherty, Department of Psychology, University of Maryland, College Park, MD 20782 (e-mail: mdougherty@psyc.umd.edu).

morrow, not rain). The focal hypothesis, *rain*, can be compared with either an implicit disjunction of *not rain* (called an implicitly packed hypothesis) or an explicit disjunction of *not rain* that consists of elements of *not rain* (e.g., snow, sleet, hail, sunshine, cloudy, all others). Tversky and Koehler argued that support was more easily generated for explicit disjunctions than for implicit disjunctions. Thus, the judged probability of the focal hypothesis tends to be larger, when compared with the implicit disjunction, because the support generated for the implicit disjunction is less than the support generated for the elements of the explicit disjunction. One consequence of failing to generate adequate support for an implicit disjunction is that judgments of implicit disjunctions will tend to be subadditive with respect to the sum of the judged probabilities of the corresponding explicit disjunctions (Fox & Tversky, 1998; Koehler, 2000; Mulford & Dawes, 1999; Tversky & Koehler, 1994). In the context of the weather example, the support for *not rain* would be less than the sum of the support for the elements of *not rain*. Thus, *p*(rain) will be lower when the support for *rain* is compared with the explicit disjunction of *not rain*.

A third model that assumes a comparison process is Dougherty et al.'s (1999) MDM model. Although MDM's comparison process is consistent with support theory, it diverges from support theory and the comparison heuristic in that it provides a mechanism by which the strength of the various hypotheses (i.e., their support) is assessed. According to MDM, the probability of the focal hypothesis is given by the ratio of the memory strength for the focal hypothesis to the sum of the memory strengths for all explicitly considered alternative hypotheses (see also Dougherty, 2001), where the memory strengths are derived using a global memory matching process.

We assume that the probability question prompts participants to generate alternative hypotheses. Returning to the weather example, if participants are asked to judge *p*(rain), it is assumed that they generate elements of *not rain*. Thus, if the decision maker generates only the *sunny* hypothesis, then $p(\text{rain, not rain}) = ms(\text{rain})/[ms(\text{rain}) + ms(\text{sunny})]$, where ms = memory strength. However, if the decision maker generates *sunny*, *sleet*, and *snow*, then $p(\text{rain, not rain}) = ms(\text{rain})/[ms(\text{rain}) + ms(\text{sunny}) + ms(\text{sleet}) + ms(\text{snow})]$. Generally, the more alternatives included in the comparison process, the lower the judged probability of the focal hypothesis. Because some hypotheses can contribute greater strength values than other hypotheses can, the judged likelihood of the focal hypothesis depends on the strength of the alternatives, not merely on the number of alternatives. Thus, the judged probability of the focal hypothesis is assumed to decrease as the overall strength of the explicitly considered alternatives increase.

Note that using the ratio of memory strengths for the focal hypothesis versus the alternatives ensures that the sum of the probabilities for all explicitly considered hypotheses is 1.0. In other words, the total probability is partitioned across all explicitly considered hypotheses (cf. Fox & Rottenstreich, 2003). For example, suppose that one judges *p*(rain, not rain), *p*(snow, not snow), *p*(sleet, not sleet), and *p*(sunny, not sunny) and always includes the three nonfocal alternatives in the comparison process [i.e., when judging *p*(rain, not rain), one considers *snow*, *sleet*, and *sunny*]. We assume that the probabilities of these four hypotheses form an implicit partition of the probability space and, therefore, are additive. We refer to this assumption as *constrained additivity* (an extension of binary complementary), because additivity is assumed to be constrained to the explicitly considered alternatives. Note that increasing the number of alternatives leads the total probability to be divided among more alternatives.

The discussion above suggests that judged probability decreases as a function of increases in the number and strength of the explicitly considered alternatives. Of importance for the present paper are the processes that govern how many and which alternatives are explicitly considered. In this article, we consider two such processes: (1) working memory (WM) processes and (2) long-term memory (LTM) retrieval. We propose that the number of alternatives included in the comparison process is intimately tied to individual differences in WM capacity and to the amount of time one has to generate relevant alternatives from LTM.

Working Memory Capacity

Recent research on WM theory has revealed that one of the primary functions of WM is attentional control and that measures of *WM capacity* reflect individual differences in participants' ability for controlled attention (Cowan, 1999; Engle, 1996; Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Bleckley, Conway, & Engle, 2001). In agreement with Baddeley's (1993; Baddeley & Hitch, 1974) conception of the central executive, Engle et al. (1999) proposed that WM capacity reflects one's ability to maintain information in the focus of attention in the face of distracting or interfering stimuli. WM capacity was distinct from short-term memory (STM) capacity in that STM capacity was assumed to reflect the relatively passive maintenance of knowledge units. Similarly, Cowan (1999) argued that the contents of WM consisted of a subset of STM—the part of STM in the focus of attention. Measures of WM span have been proposed to capture one's ability for sustainable controlled attention—a process necessary for both *maintaining* goal- or task-*relevant* information in the focus of attention and *inhibiting* goal- or task-*irrelevant* information from entering the focus of attention (Conway, Cowan, & Bunting, 2001; Engle, 2002; Kane et al., 2001).

We assume that WM capacity is fundamental to the accuracy of probability judgments, inasmuch as one's capacity should determine the number of relevant alternative hypotheses that can be maintained in the focus of attention and included in the comparison process. Assuming that participants compare the focal hypothesis with relevant alternatives, we hypothesized that the number of alternatives with which the focal hypothesis is compared is constrained by one's WM capacity—one's ability to maintain alternative hypotheses in the focus of attention while forming a probability judgment.

Although no research has directly examined the role of WM capacity in probability judgment, several studies have suggested that it is important. Mehle (1982) noted that auto mechanics tended to consider from four to six total hypotheses, or causes of auto failure. Elstein, Shulman, and Sprafka (1978) observed that expert physicians tended to think of approximately four alternative diagnoses prior to settling upon one. Finally, Dougherty, Gettys, and Thomas (1997) revealed that participants spontaneously considered only around two alternative causal scenarios when asked to judge the likelihood of a particular causal scenario. In all cases, the total number of hypotheses participants considered approximated 4 ± 1, the number of chunks that can be reliably maintained in WM (Cowan, 2001). We argue that these findings are not coincidental but actually reflect WM limitations.

Recall and Time Constraints

A second variable that should affect the number of alternatives included in the comparison process is the amount of time available to generate alternatives. We assume that the generation process involves recall from LTM and that this process takes time to complete. The more time one has to generate alternatives, the more alternatives one should be able to retrieve. Thus, participants should be able to retrieve relatively more alternatives in the absence of time constraints than when time is limited, because time constraints should truncate the memory search process. Assuming that judged probability is based on the comparison of the strength of the focal hypothesis with the strength of the generated alternatives, judged probability should be higher, and more subadditive, when retrieval time is constrained.

Overview of Experiments

The purpose of the present research was threefold. First, we sought to examine the extent to which the tendency for judgments to be subadditive was affected by the strength of the alternatives with which the focal hypothesis was compared. According to all three theories reviewed above, the judged probability of the focal hypothesis should decrease as the strength (or support) of the explicitly considered alternatives increases. In Experiment 1, we manipulated the objective frequency of the items within each distribution to alter the strength of the alternatives. Thus, overall subadditivity should decrease to the extent that the judged probabilities of the individual hypotheses in an explicit disjunction are disproportionately affected by the strongest or most likely alternatives (cf. Windschitl & Wells, 1998).

A second but related hypothesis concerns the effect of placing time constraints on the hypothesis generation process. We assume that time constraints lead participants to consider fewer alternatives when judging the focal hypothesis. In the context of the comparison process, participants should provide higher probability judgments when the generation process yields fewer alternatives. Thus, in Experiment 2, we manipulated the amount of time participants had to generate alternatives.

A third purpose of our research was to examine the relationship between WM capacity and probability judgment. We assumed that the number of alternatives with which the focal hypothesis can be compared is limited by one's WM span—one's ability to maintain the relevant alternatives in the focus of attention. Participants who can maintain, and compare, several alternatives with the focal hypothesis should give lower probability judgments and be less subadditive than participants who can maintain, and compare, only one or a few alternatives to the focal hypothesis (cf. Dougherty et al., 1997). Thus, participants high in WM span should give lower probability judgments and be less subadditive than participants low in WM span, because high WM span individuals are presumed to compare the focal hypothesis with relatively more alternatives than do low-span participants. Moreover, if attentional control is the central construct underlying the predicted negative correlation between WM span and judgment, the relationship between WM span and judgment should remain significant even after variance due to STM span is removed.

Our predictions can be derived straightforwardly within the context of an experiment. In Experiment 1, participants viewed a simulated restaurant task in which they learned how often each of four patrons ordered particular menu items from each of four different menus, with each customer ordering from one and only one menu throughout the course of the experiment. Each menu had associated with it a particular distribution with which the various menu items were ordered. For example, in the 20–30–14–2–2–2–2–2 distribution, one item was ordered 20 times, one item 30 times, one item 14 times and so on. In the 20–10–9–9–8–8–8–2 distribution, one item was ordered 20 times, one ordered 10 times, two were ordered 9 times, and so on. The other two distributions were 20–20–20–3–3–3–3–2 and 20–16–15–14–3–2–2–2. Thus, over the course of the experiment, each customer ordered the same number of menu items, but with different relative frequencies. After viewing the restaurant simulation, the participants rated the likelihood of all 32 menu items from the four distributions. In an attempt to measure hypothesis generation, the participants performed a thought-listing task (Cacioppo & Petty, 1981) for one item from each distribution. Thought listing is a form of retrospective protocol analysis in which participants are asked to list those thoughts they had while performing a task. In our case, the participants were instructed to list alternatives they thought of when making their probability judgments.

Under the normative model, the sum of the eight probability judgments for each menu should be 1.0 (or 100%), because the eight items constitute a complete partitioning of the sample space. However, as was reviewed above, considerable research has shown that judgments are often subadditive, summing to greater than the total possible probability (in this case 1.0). Thus, we expected that judgments would be subadditive. Importantly, we predicted that the degree to which judgments would be subadditive would be affected both by the distribution of the alternatives and by the participants' WM spans.

The predictions regarding the effect of distribution can be derived using the objective frequencies as a measure of strength and assuming that participants make each focal judgment by comparing its strength with the strength of a subset of alternatives in the distribution. Assume that participants consider only the two strongest alternatives within each distribution when judging each focal hypothesis. For example, in the 20–30–14–2–2–2–2–2 distribution, the likelihood of the item that occurred 30 times would be made by comparing it with the items that occurred 20 and 14 times [p(item 30) = 30/(30 + 20 + 14) = .47]. Likewise, when participants judge the likelihood of an item that occurred 2 times, the judgment would be based on the comparison between it and the two strongest alternatives [2/(30 + 20 + 2) = .04]. Assuming that all eight judgments within each distribution are made similarly, subadditivity would be greatest for the 20–10–9–9–8–8–8–2 (sum of the probabilities = 1.92) distribution and least for the 20–30–14–2–2–2–2–2 (sum of the probabilities = 1.19).[1]

In addition to examining overall subadditivity, we were also interested in examining participants' judgments for the items that occurred 20 times in each distribution. Windschitl et al. (2002) revealed that participants tended to give higher probability judgments when the focal events (i.e., items with a frequency of 20) are embedded in the context of weak alternatives. Thus, in addition to expecting an effect of distribution on subadditivity, we expected an effect of distribution on judgments of the items with a frequency of 20.

The contribution of WM span can be analyzed two ways. First, we can examine the correlation between WM span and subadditivity. As was previously stated, we predicted a negative correlation between WM span and subadditivity. However, there are two reasons that a negative correlation might manifest. First, WM span might limit the number of alternatives used in the comparison process. If this were the case, the distribution of the alternatives should affect subadditivity regardless of one's WM span. Thus, isolation of participants high and low on WM span should reveal that both groups are affected by distribution. Second, WM span might be correlated with crystallized intelligence, which might be related to probability judgment. That is, it is possible that participants with high WM span scores also have knowledge that judgments should be additive and, therefore, might try to make their judgments additive. If this were the case, distribution should have no effect (or at least a smaller effect) among high-span participants. If high-span participants try to force their judgments to be additive, the strength of the alternatives should be irrelevant. Thus, isolation of participants high in WM span should reveal no effect of distribution (or equivalently, there should be a WM span × distribution interaction). In addition, we might expect judgments from high WM span participants to approximate additivity.

There are two reasons a WM span × distribution interaction might obtain. The first reason was alluded to above. WM span might be correlated with crystallized intelligence, which might lead high-span participants to utilize their knowledge of probability theory to force their judgments to be additive. If this were the case, distribution should have a smaller effect (or no effect) among high-span participants than among low-span participants, because high-span participants should be additive regardless of distribution. Thus, the failure to find an interaction would rule out the crystallized intelligence hypothesis.

The second reason that WM span might interact with distribution has to do with the number of alternatives used in the comparison process. Increasing the number of alternatives in the comparison process might have a larger effect for the even distributions (e.g., 20–10–10–9–9–8–8–8) than for the uneven distributions (e.g., 20–30–14–2–2–2–2–2). For example, assume that participants always include the most likely alternatives in the comparison process, that the number of alternatives generated is the same for each distribution, and that low-span participants generate three alternatives and high-span participants five alternatives. These calculations yield an interaction between WM span and distribution, with high-span participants predicted to have subadditivity scores of 1.09 and 1.42 and low-span participants to have scores of 1.19 and 2.06 for the 20–30–14–2–2–2–2–2 and 20–10–10–9–9–8–8–8 distributions, respectively.

One aspect of the experimental design that might undercut such an interaction is that the strength of the alternatives is confounded with the accessibility of the alternatives. Therefore, it is unclear whether participants would generate the same number of alternatives for each distribution, especially given that the distributions contain alternatives that vary in strength. For example, items that occurred only twice in the study phase would be less accessible than items that occurred eight times. Thus, it is possible that participants would think of more alternatives for the 20–10–10–9–9–8–8–8 distribution than for the 20–30–14–2–2–2–2–2 distribution. To the degree this is the case, any interaction between distribution and WM span would be mitigated.

To summarize, we hypothesized greater probability judgment and greater subadditivity for the distributions that contained primarily weak alternatives (e.g., 20–10–10–9–9–8–8–8). Second, we predicted that placing time constraints on the generation process would lead to fewer alternatives being included in the comparison process and, consequently, to greater probability judgments and greater subadditivity than when no time constraints were present. Third, we predicted that probability judgments and subadditivity would be correlated with a measure of

WM capacity. Experiment 1 tested the first and third hypotheses, using a learning task. Experiment 2 tested the second and third hypotheses, using a state population judgment task.

EXPERIMENT 1

Method

Participants

The participants were 64 individuals from the University of Maryland campus community, who responded to an advertisement posted at various locations around campus. The participants received $10 for partaking in the experiment.

Procedure

Each participant completed the entire experiment, including the collection of individual differences data, individually in sessions that lasted between 1 and 1¼ h.

Measures of WM span and STM span. Prior to beginning the experimental task, each participant completed the operation span (o-span) task as a measure of WM span and a simple word span task as a measure of STM span (Conway & Engle, 1996; Turner & Engle, 1989). The o-span task required the participants to retain a list of words while solving mathematical problems. For example, on successive presentations, participants would be shown $(4*3) - 3 = 9$? door, $(4/2) + 3 = 7$? shoe, and so on. The participants were required to read aloud the equation, verify whether the equation was correct, and then read aloud the word. After saying the word, the experimenter advanced 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. The participants were presented with 15 sets of equation–word pairs, with set sizes ranging from two to six. Each set size occurred three times in random order. Performance on the o-span task was given by the number of words recalled in the correct serial position. The maximum possible score was 60 if the participants correctly recalled all the words from the 15 lists perfectly. The simple word span task was similar to the o-span task, with the following exceptions: (1) There was no secondary task (i.e., no math task) performed simultaneously with remembering the list of words, (2) the lists of words ranged from two to seven items, and (3) the maximum number of words that could be recalled was 81 (three blocks each of list lengths of 2, 3, 4, 5, 6, and 7). Detailed descriptions of the o-span and simple word span tasks are presented in Turner and Engle.

Practice session. Prior to engaging in the experimental task, the participants completed a practice session that resembled the experimental task. The practice session involved learning and then estimating the likelihood of where two travelers went for vacation. The participants were instructed to judge the likelihood of each focal hypothesis (e.g., vacation destination) by thinking of relevant alternatives. To encourage this, the participants engaged in a thought-listing task in which they were to type into the computer the alternatives they considered when making the judgment. This thought-listing task was also implemented in our experimental task, but we manipulated whether the participants engaged in the thought-listing task before or after making their probability judgments (see below).

Experimental task. The experimental design was a 4 (distribution of alternatives) × 2 (location of thought-listing task) mixed factorial, with distribution manipulated within subjects and location of the thought-listing task manipulated between subjects. The experimental task involved a simulation of "days" in a restaurant. The task included four menus with 8 items on each menu (total of 32 menu items). These menus were breakfast (bagel, Cheerios, fried eggs, French toast, muffin, omelet, pancakes, waffle), snack (almonds, apple, candy, Chex mix, popcorn, pretzels, French fries, Hershey's kiss), dinner (burger, chili, pasta, soup, steak, stir fry, turkey, pizza), and dessert (apple pie, banana split, brownie, cake, coffee, ice cream, pie, sundae). One simulated customer was associated with each of the four menus (e.g., only Bob ordered from the breakfast menu). Each day, each customer ordered one of the 8 menu items from the respective menu. The choice of menu item 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 20–30–14–2–2–2–2–2 (breakfast), 20–20–20–3–3–3–3–2 (snack), 20–16–15–14–3–2–2–2 (dessert), and 20–10–9–9–8–8–8–2 (dinner). Items within each menu were counterbalanced across relative frequencies, such that each item within each menu occurred at each objective frequency for an equal number of participants.

The participants proceeded through a series of "days," during which they saw each of the regular customers and what they ordered. Each "day" proceeded in the following manner: (1) an 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," and (5) a picture of a sunset that signified the end of the day. The participants saw 296 menu items over 74 days, with each menu item being presented for 3 sec. To ensure that the participants were attending to the task, we prompted the participants to recall the most recent item ordered by a particular customer 12 times during the learning phase, with each customer being used as the memory test item 3 times.

The probability judgment task followed the learning phase. Prior to judging the full set of 32 items, half of the participants (chosen at random) were presented with one menu item from each distribution that had occurred 20 times during study. These four items were presented individually one at a time on the computer, and the order was randomized for each participant. For each of these four menu items, the participants were asked to (1) judge the probability that the customer would order that menu item on the 75th day and (2) list those alternatives they thought of while they were making their probability judgment (this is referred to as the thought-listing task). Tversky and Koehler (1994) proposed that participants were less likely to generate alternatives without explicit instructions to do so. Hence, this thought-listing task served to encourage the participants to generate alternatives to compare with the focal hypothesis. The remaining half of the participants were given the thought-listing task at the end of the experiment. The generate-before condition was included because we wanted to ensure that the participants were sufficiently motivated, or encouraged, to think of alternatives in the restaurant task. The generate-after condition was included to examine whether having the participants generate in the restaurant task per se had an effect on the degree to which they were subadditive. The location of the generation was a second independent variable in our design, although we did not anticipate an effect of this manipulation. Of note, we reasoned that the thought-listing procedure might provide a weak measure of the participants' ability to generate or think of alternative hypotheses and that this might correlate with WM span. Also of note, examination of the thought-listing data revealed that the participants listed only items that were on the respective menus. For example, if the participants were judging a breakfast item, they listed only other breakfast items in their thought listings.

Immediately following either the learning phase (generate-after condition) or the generation task (generate-before condition), the participants judged the probability of each of the 32 menu items 1 at a time in random order. These 32 judgments (8 per meal) enabled us to assess the degree to which the participants' judgments were subadditive—the degree to which the sum of the probability judgments for each set of 8 menu items deviated from 100. Under the normative model, the sum of the probabilities for each set of 8 menu items should be 100, since the 8 items from each menu formed a complete partitioning of the sample space. The scale used for all the judgments was an 11-point scale (*0%–impossible*, *10%*, *20%*, *30%*, *40%*, *50%*, *60%*, *70%*, *80%*, *90%*, or *100%–certain*). Although this scale is expressed in units of 10%, it is common for participants to respond in increments of 5% or 10% even when given an open-response scale (Budescu, Weinberg, & Wallsten, 1988; Wallsten, Budescu, & Zwick, 1993). Care was taken to point out that the judgments should be made by considering how often the items had occurred throughout the entire experiment, not what had been ordered most recently. This instruction was emphasized during the practice session, as well as at the beginning of the judgment phase in the restaurant task.

Results and Discussion

Table 1 shows the mean sum of the probability judgments for the eight menu items from each distribution for both the before and the after conditions. The degree to which judgments were subadditive can be calculated by subtracting 100 from the means. Individual *t* tests revealed that the mean sum of the eight judgments was significantly greater than 100 for all four distributions (all ps $< .0001$), indicating that the judgments were subadditive. In fact, an examination of individual participants' judgments (averaged across distribution) revealed that 63 of 64 participants (98%) demonstrated subadditivity.

Distribution of alternatives. We hypothesized that the degree of subadditivity would be sensitive to the distribution of the alternatives. Consistent with this hypothesis, there was a significant main effect of distribution on the degree to which the judgments were subadditive [$F(3,60) = 15.88$, $MS_e = 1{,}175.7$, $p < .0001$]. The participants' judgments were more subadditive when the distribution included a few highly likely alternatives (20–30–14–2–2–2–2–2) than when the alternatives were all relatively unlikely (see Table 1). The effect of the location of the generation task was nonsignificant [$F(1,62) = 0.30$, $MS_e = 40{,}931$, $p > .20$], although there was a significant location × distribution interaction [$F(3,60) = 3.51$, $MS_e = 1{,}175$, $p < .05$]. An analysis of the simple effects revealed that the effect of distribution was significant for both the before [$F(3,29) = 12.22$, $MS_e = 1{,}393.1$, $p < .001$] and the after [$F(3,29) = 5.07$, $MS_e = 958.4$, $p < .01$] conditions. The means for both the before and the after conditions showed similar patterns: Subadditivity was lowest for the two distributions that contained the strongest alternatives (20–30–14–2–2–2–2–2 and 20–20–20–3–3–3–3–2) and highest for the two distributions that contained relatively weak alternatives (20–16–15–14–3–2–2–2 and 20–10–10–9–9–8–8–8). Post hoc analyses on the between-subjects variable (before vs. after) failed to reveal any significant differences for any of the four distributions, even using a liberal alpha of .10 for each of the four comparisons. Given that none of these pairwise comparisons were significant, we are hesitant to interpret the location × distribution interaction. Collapsing across before and after conditions, planned comparisons using Bonferroni adjusted *t*s (alpha set at .008 per comparison) revealed that all pairwise comparisons were significant, except for the contrast between the 20–30–14–2–2–2–2–2 and the 20–20–20–3–3–3–3–2 distributions. These results clearly show that subadditivity was affected by the distribution of alternatives.

There were no significant effects of distribution or location of the thought-listing task on the number of alternatives the participants listed in the thought-listing task, nor did these variables affect judgments of the items presented with a frequency of 20 (freq-20 and freq-20TL in Table 2). The latter finding was surprising, especially given that Windschitl et al. (2002) showed an effect of distribution in a similar learning-based paradigm. We discuss possible reasons for the failure to find such an effect in the General Discussion section.

Recall that the participants judged one freq-20 item from each distribution on two occasions: once in the context of the thought-listing task (which took place either before or after judging the exhaustive set of 32 items) and once in the context of judging the 32 items. One might raise two questions. First, were judgments of the freq-20 items affected by the thought-listing task? Second, did having performed the thought-listing task affect the participants' judgments that were made subsequently in the context of judging the exhaustive list of 32 items (a carryover effect of sorts)? That is, were judgments of the items made while judging the exhaustive set influenced by whether or not the thought-listing task came before or after rating the exhaustive set?

The first of these questions can be answered by examining whether judgments for the freq-20 items were significantly different for judgements made in the two different contexts. The second can be answered by examining whether the context of the freq-20 judgments (during thought listing vs. during the exhaustive set) interacted with location (before or after) of the thought-listing task.

Table 1
Mean Sum of the Probability Judgments (With Standard Errors) for the Eight Items Within Each Distribution for the Generate-Before and Generate-After Conditions in Experiment 1

	Distribution							
Generate	20–30–14–2–2–2–2–2		20–20–20–3–3–3–3–2		20–16–15–14–3–2–2–2		20–10–9–9–8–8–8–2	
Condition	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Before	243.4	16.4	249.4	17.7	265.3	19.3	307.5	21.3
After	245.0	17.3	240.0	17.8	255.9	18.4	269.4	20.5
Overall mean	244.2^{a}	11.8	244.7^{a}	12.5	260.6^{b}	13.2	288.4^{c}	14.8

Note—Distributions with different superscripts in the overall mean row are statistically different from one another by Bonferroni adjusted *t* ($\alpha = .008$ per comparison).

Table 2
Mean Judgments (With Standard Errors) for the Freq-20 and Freq-20TL Items in Experiment 1

	Distribution							
	20–30–14–2–2–2–2–2		20–20–20–3–3–3–3–2		20–16–15–14–3–2–2–2		20–10–9–9–8–8–8–2	
Items	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Freq-20	55.4	2.8	52.8	2.9	53.1	2.7	53.7	2.9
Freq-20TL	50.3	2.7	50.6	2.8	48.9	2.6	48.4	2.7

Note—Freq-20, items presented with a frequency of 20 during the probability judgment phase; Freq-20TL, items presented with a frequency of 20 during the thought-listing phase.

Collapsing across distributions, we found that judgments of the freq-20 items were significantly lower when they were made as part of the thought-listing task ($M = 49.5$) than when they were made as part of rating all 32 menu items [$M = 53.8$; $F(1,62) = 25.46$, $MS_e = 22.3$, $p < .0001$]. Note, however, that the participants overestimated the objective probability of the focal (.40) in both cases. Interestingly, this did not interact with the location of the thought-listing task [before or after rating the 32 items; $F(1,62) = 0.55$, $MS_e = 691.0$, $p > .10$]. This suggests that having the participants engage in the thought-listing task at the beginning of the experiment did not affect how they rated the same freq-20 items when they were judged in the context of all 32 menu items. However, it also suggests that the thought-listing task affected how the participants rated the hypotheses that they judged as part of the thought-listing task. Judgments were lower when made in the context of the thought-listing task.

WM span and probability judgment. Our second set of hypotheses concerned the relationship between WM span and subadditivity, and between WM span and freq-20 items. Table 3 presents the zero-order correlations between WM span, STM span, subadditivity, and freq-20 made in the context of judging the exhaustive set, and the freq-20 items made in the context of the thought-listing task (freq-20TL).[2] The number of alternatives listed in the thought-listing procedure did not correlate with WM span, STM span, subadditivity, or freq-20 judgments. However, as can be seen, both WM span and STM span correlated significantly with the sum of the probability judgments and the freq-20 judgments.

We theorized that WM span, not STM span, was the crucial component underlying individual differences in probability judgment. If this is true, WM span should remain a significant predictor even after variance due to STM span is partialled out. Table 4 presents these partial correlations. The relationship between WM span and the sum of the probability judgments and between WM span and freq-20TL judgments remained significant even after controlling for variance due to STM span. In contrast, the correlations between STM span and subadditivity ($r = -.15$, $p = .22$) and STM span and freq-20 judgments ($r = -.10$, $p = .42$) with WM span partialled out were quite small and nonsignificant.

To get a better sense of the relationship between WM span and subadditivity, we compared the mean sum of the probability judgments for participants scoring in the highest and lowest thirds on the WM span task (see Figure 1). Collapsing across distribution, the mean sum of the probabilities for high-span participants ($M = 218.8$) was nearly 100 points lower than that for low-span participants [$M = 311.5$; $t(44) = 3.18$, $p < .01$]. Effect sizes calculated using Cohen's *d* revealed that this difference was quite large ($d = .93$), indicating that nearly a full standard deviation separated the high- and the low-span participants. The pattern was similar for freq-20 judgments, with high-span participants ($M = 47.3$) providing lower estimates than did low-span participants [$M = 59.0$; $t(44) = 2.04$, $p < .05$; $d = .68$]. Together with the finding that WM span is correlated with judgments even after partialling out STM span, this suggests that the crucial factor was not merely the number of alternatives the participants could passively maintain but, rather, their ability to maintain alternatives in the focus of attention while comparing them with the focal hypothesis.

One possible criticism of our analysis is that it is correlational. Thus, it is possible that the relationship between WM span and subadditivity is mediated by crystallized intelligence. Although our experiment does not

Table 3
Zero-Order Correlations for Main Dependent Variables in Experiment 1

	STM Span	WM Span	Freq-20	Freq-20TL
WM span	.556****			
Freq-20	−.33**	−.32*		
Freq-20TL	−.29*	−.37**	.94****	
Sum of probability judgments	−.35**	−.42***	.89****	.88****

Note—Freq-20, items presented with a frequency of 20 during the probability judgment phase; Freq-20TL, items presented with a frequency of 20 during the thought-listing phase; STM, short-term memory; WM, working memory. *$p < .05$. **$p < .01$. ***$p < .001$. ****$p < .0001$.

Table 4
Partial Correlations Controlling for Short-Term Memory Span in Experiment 1

	WM Span	Freq-20	Freq-20TL
Freq-20	−.17		
Freq-20TL	−.26*	.93****	
Sum of probability judgments	−.29*	.88****	.87****

Note—Freq-20, items presented with a frequency of 20 during the probability judgment task; Freq-20TL, items presented with a frequency of 20 during the thought-listing task; WM, working memory. $^*p < .05$. $^{****}p < .0001$.

completely rule out this possibility, there are two aspects of the data that argue against this interpretation. First, if the participants high in WM span had knowledge that judgments should sum to 100, presumably there should be several participants who had additive judgments for at least one of the four distributions. However, additivity held only twice out of the 256 chances (64 participants × 4 distributions = 256), and only 1 participant showed superadditivity (judgments summing to less than 100). Thus, any effect of knowledge on additivity was minimal, since few participants showed overt signs of being able to use their knowledge to produce additive judgments.

A second aspect of the data that can be examined is whether the high- and the low-span participants were equivalently affected by the distribution. Note that if the high-span participants were trying to force their judgments to be additive, there should be no effect of distribution on their judgments or, minimally, a much smaller effect of distribution. In contrast, if the participants used a comparison process to make their judgments but the high-span participants included more alternatives in the comparison, both the high- and the low-span participants should be affected by distribution. Thus, if the low-span participants were affected by distribution but the high-span participants were not (or were affected to a lesser degree), this would open up the possibility that part or all of the correlation between WM span and subadditivity was due to differences in knowledge of probability theory. The failure to find such an interaction would lend support for the idea that WM capacity constrains the number of alternatives used in the comparison process.

Comparisons of the effect of distribution on participants' scoring in the upper and lower thirds (see Figure 1) revealed a main effect of distribution [$F(3,37) = 8.35$, $MS_e = 1,297.8$, $p < .01$] and a main effect of WM span [$F(1,39) = 11.37$, $MS_e = 39,878.1$, $p < .01$] on subadditivity, with no distribution × WM span interaction [$F(3,37) = 0.07$, $MS_e = 1,297.8$, $p = .97$]. Separate analyses on low- and high-span groups confirmed that the effect of distribution was significant for both low-span [$F(3,18) = 6.58$, $MS_e = 1,485.2$, $p = .015$] and high-span [$F(3,17) = 3.31$, $MS_e = 1,100.5$, $p = .04$] participants. Thus, distribution had an equally large effect on subadditivity regardless of the participants' WM span. In sum, that in only 2 out of 256 chances did judgments sum to 100, together with the finding that distribution affected high-span and low-span participants equally, suggests that knowledge that judgments should be additive was not a major factor in our results.

One question of interest is whether WM span mediates the effect of distribution. That is, are the variance accounted for by our independent variable and the variance accounted for by WM span independent? One way to test this is to examine the effect of distribution with and without WM span in the model and examine whether the effect of distribution increases or decreases when WM span is in the model, relative to when it is not in the model. If the effect of distribution is mediated by WM

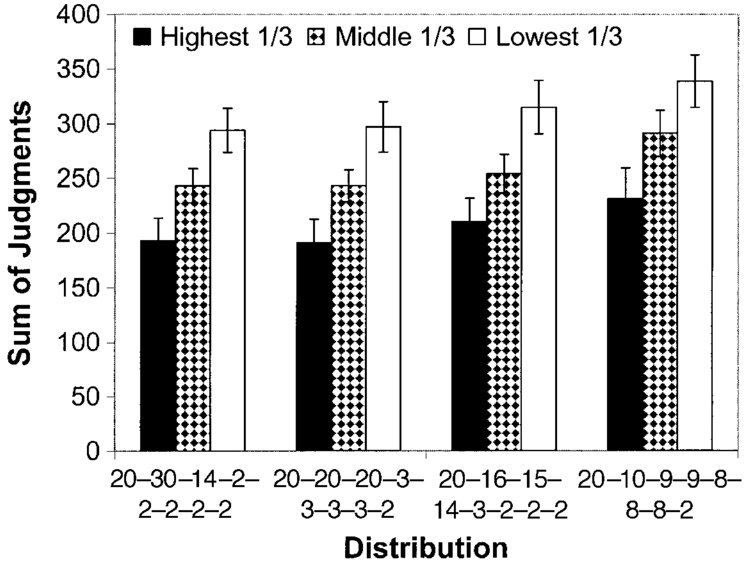

Figure 1. Mean sum of the probability judgments for participants scoring in the upper, middle, and lower thirds on the operation span task in Experiment 1.

span, the amount of variance due to distribution (η^2) when WM span is added to the model should decrease. In contrast, if WM span accounts for a significant amount of unique variance (i.e., reduces random error), the η^2 due to distribution should actually increase when WM span is included as a predictor, because inclusion of the predictor would reduce the random error variance. Examination of the effect of distribution yielded $F(3,60) = 15.89, p < .0001$, and $\eta^2 = .36$ when span was not in the model and $F(3,59) = 2.83, p = .04$, and $\eta^2 = .06$ when span was included in the model. Clearly, the effect of distribution was reduced when WM span was included in the model, suggesting that WM span mediated the effect of distribution.

EXPERIMENT 2

The purpose of Experiment 2 was twofold. First, we wanted to extend our basic findings in Experiment 1 to a novel task in order to explore the generality of the WM-span–subadditivity relationship. Thus, rather than using the restaurant task, we used a state population judgment task in which participants rated the likelihood that a randomly selected individual lived in a particular state. The second purpose was to explore the effect of constraining how much time participants had to retrieve alternatives on subadditivity. We theorized that the hypothesis generation process requires the retrieval of alternatives from LTM that are then used in the comparison process. Assuming that hypothesis generation (i.e., recall from LTM) takes time and that placing time constraints on the judgment process truncates the generation process, we hypothesized that the participants would be more subadditive under time constraints than under no time constraints. In addition to addressing these two empirical questions, we also changed how the participants responded with their probability judgments. Rather than having the participants provide judgments in increments of .10 by clicking on a button, we had them type their responses, using the keyboard. This enabled the participants to use the entire probability scale.

Method

Participants

Ninety-five participants completed the experiment. The participants were recruited through advertisements posted around the University of Maryland campus and were paid $13 for their participation.

Design

A 2 (time constraints: present or absent) × 2 (counterbalance order: time constraints first or last) mixed factorial design was used, with the presence or absence of time constraints as the within-subjects variable and counterbalance order as the between-subjects variable. Measures of WM capacity, STM capacity, and verbal fluency were included to serve as predictor variables.

Procedure

Prior to the experiment, the participants were shown the map of the entire United States partitioned into the regions defined by the United States Census Bureau, shown in Figure 2. For each trial, the participants were shown each state in the context of the appropriate region and were asked to judge the probability that a randomly chosen person lived in a particular state, given the knowledge that they came from that region. For example, “Imagine that I choose, at random, one person who lives in the Western region. What is the probability that this person lives in California, given that he or she is from the Western region?” The participants rated all the states from all four regions. For half of the regions, judgments of state populations were made under time constraints. The participants were given 6 sec to make these judgments. This time was set on the basis of an estimate that was one standard deviation below the mean time it took participants to make judgments in a pilot study using the same task (see Svenson & Benson, 1993). In order to prepare the participants for the speeded judgment task, they engaged in a speeded practice session immediately prior to the judgment task. The practice session consisted of 10 anagrams, each of which had to be solved within 6 sec. Judgments that took longer than 6 sec were followed by feedback saying “Please respond faster!” Then instructions concerning the population judgment task were briefly repeated, emphasizing both speed and accuracy. Finally, the participants were given states from two regions intermixed and were asked to type a number between 0 and 100 indicating the probability that a randomly selected individual from the region would reside in each state. Each state was shown on an individual screen, and the computer generated a random order of the states for each participant.

For the other half of the state judgments, the participants were given no time limit and were encouraged to take as much time as needed to make their judgments. The practice session for the nonspeeded judgment task encouraged careful thought, not speed. Again, the participants were presented with 10 anagrams prior to

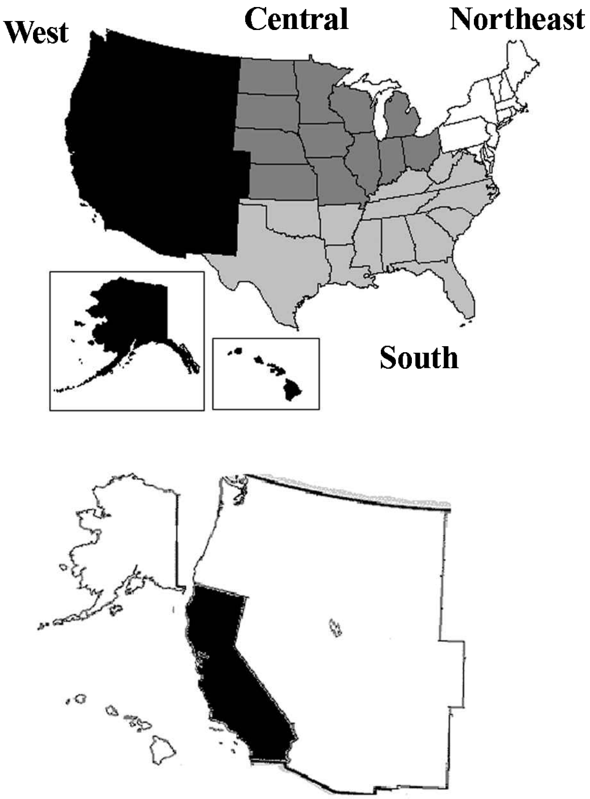

Figure 2. Top: Map of U.S. Census divisions used in Experiment 2. Bottom: Example stimulus used in the state judgment task, in which the participants are asked to judge *P*(California | West).

making their judgments, but they were encouraged to give as many solutions as they could without a time limit. The instructions for the state judgments were briefly reviewed, stressing accuracy instead of speed. Then, the participants judged state populations from two regions, in the same format as that described above.

The order of the time constraint versus no time constraint conditions was counterbalanced so that half of the participants received time constraints first and the other half no time constraints first. Within these conditions, the regions (South, West, Central, and Northeast) were counterbalanced so that each region occurred in the speeded and the nonspeeded condition for an equal number of participants. The participants were randomly assigned to counterbalancing condition.

Individual differences measures. Prior to the probability judgment task, individual differences data on WM span, STM span, and verbal fluency were collected. WM span and STM span were assessed using the same tasks as those utilized in Experiment 1. In accord with Conway and Engle (1996) and Turner and Engle (1989), an 85% criterion was set for accuracy on the concurrent mathematical verification task. Two participants whose performance fell below this criterion were disqualified from the experiment. Verbal fluency was measured using a category fluency task similar to that used in Rosen and Engle (1997). The participants were given a category name and then were asked to name as many exemplars from that category as possible in 5 min. After each minute of the task, they were asked to draw a line under what they had generated thus far and to continue generating. The participants engaged in the category fluency task twice, once for the category of *animals* and once for the category of *occupations*. The total number of unique exemplars generated for each category was counted. A composite fluency score was computed for each participant by converting their scores for each category to *z* scores and averaging.

Results and Discussion

Ninety-five participants completed the study. Of these, the data from 3 participants were eliminated because they responded with the same probability for all 50 judgments (2 of these participants gave judgments of 100 for all 50 questions, and 1 participant gave judgments of 50 for all 50 questions). An additional 7 participants, whose subadditivity scores were more than three standard deviations above the group means, were identified as outliers and were excluded from the statistical analyses. Of these 7 participants, 6 provided judgments that were in the hypothesized direction and in the same direction as the overall group means. The data from these 7 participants are presented in the Appendix.

Each participant judged all four regions of the U.S.: two under time constraints and two with no time constraints. Judgments within each region were summed. The two regions within each time constraint condition were then averaged to get a mean sum of the probability judgments from each participant for the time constraint and the no time constraint conditions.

Manipulation Check

Table 5 presents the mean judgment reaction times (RTs). An analysis of variance was conducted on RT, examining the effects of order and time constraints. As was expected, there was a significant effect of time constraints on RT [$F(1,83) = 86.58$, $MS_e = 5{,}781{,}680.2$, $p < .001$]. The main effect was, however, qualified by a significant interaction between counterbalance order and time constraints [$F(1,83) = 36.37$, $MS_e = 5{,}781{,}680.2$, $p < .001$], as well as a main effect of the counterbalancing condition [$F(1,83) = 8.17$, $MS_e = 9{,}586{,}451.4$, $p < .01$]. In the condition in which the no time constraint condition came first, the RT for no time constraints exceeded all other conditions by at least 3 sec. More important, in the condition in which time constraints came first, the no time constrained judgments were made in 5.7 sec, which was below the manipulated response deadline in the time constraint condition. As can be seen in Table 5, the mean RTs for the time constraint and no time constraint conditions were 4.39 and 9.19 sec, respectively, for the first set of judgments, but were 3.52 and 5.6 sec for the second set of judgments, respectively. Thus, our manipulation of time constraints (respond within 6 sec) was successful only for the first set of judgments. Evidently, there were carryover or practice effects in the second set of judgments that led the participants to respond more quickly in the no time constraint condition when they had received the time constraint condition first.

Table 5
Mean Reaction Times in Milliseconds (With Standard Errors) for the Time Constraint and No Time Constraint Conditions by Order of Judgments

Order of	Time Constraints		No Time Constraints	
Judgments	*M*	*SE*	*M*	*SE*
First	4,393.86	173	9,191.8	681
Second	3,525.57	124	5,603.50	421

Time Constraints and Subadditivity

We predicted that subadditivity would be greatest under time constraints than under no time constraints. We also speculated that participants high in WM span might be less affected by time constraints than low-span participants would be. However, given that our time constraint manipulation was successful only for the first set of judgments, one might expect that time constraints would affect only the first set of judgments. Model comparison procedures were conducted on the sum of the probability judgments, with WM span, STM span, and the verbal fluency composite scores as predictor variables, time constraint as the within-subjects variable, and order of the time constraint manipulation as the between-subjects variable. Verbal fluency and STM span did not account for unique variance in the model and were, therefore, excluded from all the analyses.

Figure 3 presents the least-squared adjusted mean subadditivity scores for the time constraint and the no time constraint conditions for the first and last sets of judgments. Overall, there was a significant effect of time constraints on the sum of the probability judgments [$F(1,82) = 5.45$, $MS_e = 2{,}201.3$, $p < .05$], a significant time constraints × counterbalancing condition interaction [$F(1,82) = 9.02$, $MS_e = 2{,}201.3$, $p < .01$], and a

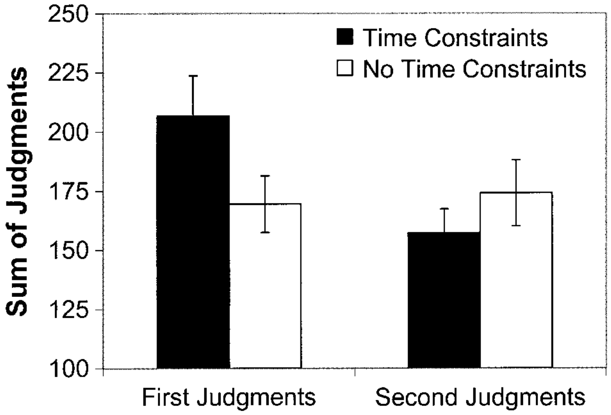

Figure 3. Mean sum of the probability judgments for the time constraint and the no time constraint conditions for the first and second blocks of judgments in Experiment 2.

time constraints × WM span interaction $[F(1,82) = 4.14$, $MS_e = 2{,}201.3$, $p < .05]$, with WM span serving as a significant negative predictor of subadditivity $[F(1,82) = 5.08$, $MS_e = 13{,}211.4$, $p < .05]$. Although the three-way counterbalance condition × time constraints × WM span interaction was nonsignificant, examination of the individual regression coefficients for the four conditions revealed that the only condition in which WM span was not a good predictor of subadditivity was under no time constraints when the no time constraint condition followed the time constraint condition. Thus, the time constraints × WM span interaction was likely due to the carryover effect for the condition that received the time constraint instructions first.

One way to examine the effect of time constraints and its relationship with WM span devoid of carryover or practice effects is to analyze only the first set of judgments. Recall that half of the participants made judgments under time constraints first, and the other half made judgments under no time constraints first. Thus, we conducted a between-groups analysis of covariance, using WM span as the covariate. This test revealed main effects of time constraints $[F(1,84) = 3.44$, $MS_e = 9{,}091.6$, $p = .06]$ and WM span $[\beta = -2.79$; $F(1,84) = 5.92$, $MS_e = 9{,}091.6$, $p < .05]$, with a no time constraints × WM span interaction ($p = .87$). This analysis suggests that high- and low-span participants were equivalently affected by the time constraint manipulation.

Correlational Analyses

Table 6 presents the zero-order correlations between the main dependent variables in Experiment 2. These correlations are consistent with Experiment 1: WM span was negatively correlated with the sum of the probability judgments. Although the correlation between WM span and the sum of the probability judgments under no time constraints was nonsignificant, it was in the same direction as that in the time constraint condition. Also of note is that finding that verbal fluency was positively correlated with WM span, but not with the sum of the probability judgments. This may suggest that WM span plays a mediating role between generation ability and judgment.

In Experiment 1, we argued that the relationship between WM span and subadditivity was driven by attentional control. This argument was based on the finding that WM span remained a significant predictor of the sum of the probability judgments after variance due to STM was removed. Table 7 presents the partial correlations controlling for STM span for Experiment 2. As in Experiment 1, the correlation between WM span and the sum of the probability judgments remained significant and, in fact, actually increased when controlling for variance due to STM span. Also of interest is that the correlation between WM span and judgments under no time constraints approached significance. These results are consistent with Experiment 1 and further support the idea that attentional control is fundamental to probability judgment.

GENERAL DISCUSSION

Our experiments revealed three important findings. First, the degree to which the participants were subadditive was affected by the distribution of the alternatives. We argue that this effect was due to a constraint on the comparison process—the participants' tendency to compare the focal hypothesis with a subset of alternatives, rather than with the exhaustive set (cf. Windschitl & Wells, 1998). Second, we showed that the degree to which the participants were subadditive was related to their ability for sustainable controlled attention, as measured by WM span tasks. As was hypothesized, the participants high in WM span were less subadditive than were the participants low in WM span. Moreover, these correlations remained significant even after variance due to STM span was removed. Third, time constraints on judgment led to increased subadditivity. We argue that these results are consistent with a model that assumes that judgments are based on a comparison of the strength of the focal hypothesis with the strength of generated alternatives. Importantly, our data indicate that the number of alternatives participants include in this comparison process is constrained by WM capacity and by the

Table 6
Zero-Order Correlations Between Main Dependent Variables in Experiment 2

	Time Constraints	No Time Constraints	WM Span	Fluency
No time constraints	.703***			
WM span	−.233*	−.144		
Fluency	−.008	.091	.205†	
STM span	−.127	−.072	.341**	.210†

Note—Time constraints, sum of probability judgments under time constraints; no time constraints, sum of probability judgments under no time constraints; WM, working memory; STM, short-term memory. *$p < .05$. **$p < .01$. ***$p < .001$. †$p = .06$.

Table 7
Partial Correlations for Main Dependent Variables When Controlling for Short-Term Memory Span in Experiment 2

	Time Constraints	No Time Constraints	WM Span
No time constraints	.681***		
WM span	−.280**	−.211†	
Fluency	−.077	.178	.125

Note—Time constraints, sum of probability judgments under time constraints; no time constraints, sum of probability judgments under no time constraints; WM, working memory. $^{**}p < .01$. $^{***}p < .001$. $^{\dagger}p = .06$.

amount of time available for generating alternatives from LTM.

One noteworthy finding was that we failed to find an effect of distribution on the judged probability of the freq-20 items, a finding that is inconsistent with recent research by Windschitl et al. (2002). One possible explanation for this failure to replicate Windschitl et al. is that our judgment task differed from their standard paradigm, in that the participants in Windschitl et al. judged the freq-20 items prior to judging any of the other items in the distributions. In our study, the participants judged the 32 alternatives in a random order or were required to engage in alternative generation when making their judgments. This may have had an effect on the nature of the comparison process by enhancing the salience of the alternatives. A second difference between our study and Windschitl et al. was that we explicitly encouraged the participants to think of alternatives with which to compare the focal hypothesis. This may have prompted the participants to consider more alternatives than they normally would have without such encouragement. Our task was also somewhat more difficult than the task used by Windschitl et al. Our experiment employed four different distributions with eight items each. Windschitl et al. used two distributions with five items each. Obviously, we cannot specify which of these possible differences led to our failure to find an effect of distribution on the freq-20 items, but it does signal a need to explore the boundary conditions of the effect.

Theoretical Implications

We began this article with a discussion of three models of probability judgment: the comparison heuristic proposed by Windschitl and Wells (1998), support theory proposed by Tversky and Koehler (1994), and MDM proposed by Dougherty et al. (1999). Although all three of these models assume a comparison process, none of these models describes the processes that underlie the comparison process. The main theoretical message of the present research is that WM processes are paramount to the comparison process and play an important role in determining the accuracy of probability judgments. More important, the present research indicates that models of WM should play a greater role in guiding theory development and research on hypothesis generation and probability judgment.

One model that is a natural candidate to be extended to account for both hypothesis generation and probability judgment is MDM (Dougherty, 2001; Dougherty et al., 1999). Although MDM assumes that judged probability is based on a comparison between the focal and the explicitly considered hypotheses (see Dougherty, 2001), the model does not explicitly account for the role of WM processes or attentional control in determining how many alternatives are explicitly considered, nor does the model include processes that enable it to generate multiple hypotheses from LTM from a single set of cues. However, as the present research clearly indicates, any adequate model of probability judgment needs to include both a limited capacity WM system and a process for generating hypotheses. Below, we will discuss an extension of MDM, a model called HyGene, that we are developing to account for the role of WM processes in hypothesis generation and probability judgment. Because HyGene is still in the development stage, our discussion of the properties of the model will be brief.

As Gettys and Fisher (1979) point out, hypothesis *generation* is the precursor to most real-world probability estimation tasks, since in most cases the hypotheses are not presented to the participant but must be generated from LTM. Consequently, we assume that the generation of hypotheses is initiated on at least two occasions: (1) when participants are prompted to evaluate a particular hypothesis and/or (2) when data from a problem are presented. We assume that data extracted from the judgment task serve as the memory cues for initiating retrieval. The retrieval goal in hypothesis generation and probability judgment tasks is to enumerate a set of relevant alternative hypotheses. Importantly, the more relevant the alternatives included in the comparison process are, the more accurate the probability judgments should be (cf. Dougherty & Hunter, 2003; Tversky & Koehler, 1994).[3] The original Minerva 2 model (Hintzman, 1988) included a mechanism for modeling recall. However, overt recall in Minerva 2 was limited to the retrieval of only a single exemplar from LTM. Thus, in order to model hypotheses generation, HyGene allows for the recall of multiple exemplars from LTM, using a single set of memory cues.

Fisher, Gettys, Manning, Mehle, and Baca (1983) showed that hypothesis generation involves a subprocess called *consistency checking*. Consistency checking is a process whereby hypotheses are checked for their logical consistency with the available data and are rejected if they are deemed inconsistent. Hypotheses deemed consistent with the data are explicitly generated and brought into WM for further processing. The consistency-checking process is assumed to consist of a rapid semantic verification process, not the explicit consideration of likelihood (Dougherty et al., 1997; Fisher et al., 1983; Gettys, Mehle, & Fisher, 1986).

The generation of inconsistent or irrelevant hypotheses is potentially problematic for two reasons. First, the more time one spends retrieving inconsistent or irrelevant hypotheses, the less time there will be that will be dedicated to retrieving relevant hypotheses. Second, irrelevant or inconsistent hypotheses that are temporarily generated may interfere with the generation of relevant hypotheses and/or take up space in WM. Thus, we assume that the consistency-checking process also involves an inhibition process, in which participants attempt to suppress the regeneration of hypotheses that were previously deemed inconsistent. Thus, once a hypothesis has been deemed inconsistent with at least a portion of the data, it should be less likely to be regenerated even if it is consistent with data that becomes available later in the task. Such inhibition processes have been shown to be fundamental to performance on verbal fluency tasks (Rosen & Engle, 1997) and in dealing with proactive interference (Kane & Engle, 2000) and have been shown to correlate highly with measures of WM span. We propose that this same type of inhibition process underlies hypothesis generation and the inhibition of irrelevant or inconsistent alternatives.

Hypotheses that survive the consistency-checking process possess the minimum requirements to be passed onto the second process, the hypothesis evaluation process. Once explicitly generated, hypotheses may be placed in the set of leading contenders (SOC). The SOC is a temporary storage space for maintaining hypotheses in WM while making a probability judgment. We assume that the number of alternatives retained in the SOC is limited by WM capacity. In most circumstances, it will be impossible for the decision maker to retain all possible alternatives in the SOC. Consequently, participants are assumed to retain only a subset of the possible alternative hypotheses; the choice of which hypotheses are retained is determined by the relative memory strengths of the hypotheses in the SOC. Thus, as hypotheses are generated from LTM and passed onto the SOC, they are evaluated for memory strength. New hypotheses that are generated replace old hypotheses only if their strength is greater than the weakest member of the SOC or if hypotheses presently in the SOC are inconsistent with newly available data.

The process of replacing members of the SOC on the basis of familiarity strength has two consequences. One consequence is that the criterion for adding new hypotheses to the SO√C becomes increasingly strict as a function of time, because the minimum strength needed to enter the SOC necessarily increases as weaker hypotheses are replaced by stronger hypotheses. In fact, Gettys and Fisher (1979) revealed this to be the case: They found that hypotheses added to the SOC later in the generation process tended to be more likely than those generated and added to the SOC earlier. The second consequence is that the hypotheses ultimately retained in the SOC will tend to be the strongest contenders to the focal hypothesis, which ultimately impacts judgment accuracy. The degree to which participants overestimate the focal hypothesis should be less when the SOC contains the strongest alternative hypotheses. In any event, although the number of hypotheses in the SOC may remain stable across time, the composition of the SOC likely will change (Gettys & Fisher, 1979).

Exactly how the generation process is terminated is uncertain. One possibility is that generation terminates when all the data have been presented. A second possibility is that termination is prompted by successive retrieval *failures*—the failure to generate new hypotheses. Finally, time constraints may force the decision maker to terminate search prematurely. In any event, once hypotheses have been generated from LTM and the SOC sufficiently populated, we assume that the likelihood judgment process is carried out using an MDM-type process, where the judged likelihood of each hypothesis in the SOC is determined by comparing its memory strength with the sum of the memory strengths of all the other hypotheses in the SOC.

Prior research on hypothesis generation, coupled with the present research, substantiates several components of the model proposed above. The present data provide evidence that the number of alternatives included in the SOC is constrained by WM limitations and that the number of alternatives used in the comparison process is affected by time constraints. Fisher et al. (1983) revealed evidence consistent with the consistency-checking process. They showed that (1) participants considered hypotheses not explicitly generated, (2) the time it took to perform consistency checking was less than the time it took to retrieve a hypothesis from LTM (suggesting that the process involves a high-speed process), and (3) participants rejected one to two hypotheses prior to explicitly generating a hypothesis. Similar findings were reported by Dougherty et al. (1997). Gettys and Fisher (1979) revealed evidence consistent with the SOC threshold. In short, Gettys and Fisher found that participants tended to become more selective in what hypotheses were included in the SOC over time. Once the SOC was filled, the minimum criterion for replacing an old hypothesis with a new hypothesis was higher. Thus, the threshold for generating a new hypothesis increased as a function of the number of hypotheses already retrieved.

Summary

In summary, the results of our experiments provide motivation for developing models of probability judgment that explicitly model the contribution of WM processes. As was mentioned in the introduction of this paper, none of the current models of probability judgment, including MDM, address how people generate or think of alternative hypotheses, nor do they address the processes responsible for maintaining those hypotheses active in memory when forming probability judgments. Instead, most models only go so far as to describe the processes that govern how people assess the likelihood of prespecified hypotheses. Interestingly, in most real-world judgment tasks, the decision maker is required to generate the to-be-judged alternatives prior to assessing their likelihood. For example, a physician must generate or retrieve from LTM a set of possible diagnoses, on the basis of the available pattern of symptoms, prior to stating their diagnosis. The present research suggests that how people generate and assess the likelihood of hypotheses is intimately tied to both LTM retrieval processes and WM limitations. Certainly, to gain a fuller understanding of the processes underlying probability judgment, theories of judgment that explicitly model these components will need to be developed.

REFERENCES

Baddeley, A. D. (1993). Working memory or working attention? In A. D. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness, and control: A tribute to Donald Broadbent* (pp. 152-170). New York: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47-89). New York: Academic Press.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception & Performance*, **14**, 281-294.

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

Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, **8**, 331-335.

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. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62-101). New York: Cambridge University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral & 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 & Human Decision Processes*, **70**, 135-148.

Dougherty, M. R. P., & Hunter, J. E. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, **113**, 263-282.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Engle, R. W. (1996). Working memory and retrieval: An inhibition-resource approach. In J. T. E. Richardson, R. W. Engle, L. Hasher, R. H. Logie, E. R. Stoltzfus, & R. T. Zacks (Eds.) *Working memory and human cognition* (pp. 89-119). New York: Oxford University Press.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, **11**, 19-23.

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 hypothesis generation. *Organizational Behavior & Human Performance*, **31**, 233-254.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, **14**, 195-200.

Fox, C. R., & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science*, **44**, 879-895.

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

Gettys, C. F., Mehle, T., & Fisher, S. (1986). Plausibility assessments in hypothesis generation. *Organizational Behavior & Human Decision Processes*, **37**, 14-33.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, **96**, 528-551.

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.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 336-358.

Koehler, D. J. (2000). Probability judgment in three-category classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 28-52.

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

Mulford, M., & Dawes, R. M. (1999). Subadditivity in memory for personal events. *Psychological Science*, **10**, 47-51.

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

Svenson, O., & Benson, L., III (1993). Framing and time pressure in decision making. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making* (pp. 133-144). New York: Plenum.

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

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

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

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

Windschitl, P. D., & Young, M. E. (2001). The influence of alternative outcomes on gut-level perceptions of certainty. *Organizational Behavior & 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 & Cognition*, **30**, 469-477.

NOTES

1. These predictions can be derived by computing what the probability would be for each focal hypothesis if it were compared only with the two strongest alternatives and then summing over the eight predicted probabilities within each distribution. For the 20–30–14–2–2–2–2–2 distribution, the summed probability would be given by $20/(20 + 30 + 14) + 30/(20 + 30 + 14) + 14/(20 + 30 + 14) + 5(2/[20 + 30 + 2]) = 1.19$. The corresponding sum of the probabilities for the remaining three distributions would be 1.33, 1.51, and 1.92 for the 20–20–20–3–3–3–3–2, 20–16–15–14–3–2–2–2, and 20–10–9–9–8–8–8–2 distributions, respectively.

2. These correlations were done by collapsing across distribution and thought-listing condition.

3. This is readily seen by inspection of Equation 1 in the context of support theory. The more alternatives unpacked from B, the lower the overall probability of A.

APPENDIX
Data From the 7 Participants Identified as Outliers

		Mean Sum of Judgments	
Participant No.	Working Memory Span Score	Time Constraints	No Time Constraints
102	23	936.5*	915.5*
105	38	945.0*	790.0*
151	21	916.0*	794.5*
153	45	818.5*	398.5
160	30	874.5*	764.0*
177	29	668.0*	231.0
179	34	512.5	695.0*
Mean		810.14	655.50

Note—Participants were excluded from analysis if one score was more than three standard deviations above the condition means. Note that the mean sums of the judgments for the time constraint and no time constraint conditions for all outliers except 179 were in the hypothesized direction. *Sum was more than three standard deviations above the group mean.

(Manuscript received May 2, 2002;
revision accepted for publication April 13, 2003.)