frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 17 June 2011
doi: 10.3389/fpsyg.2011.00129

Implications of cognitive load for hypothesis generation and probability judgment

Amber M. Sprenger[1], Michael R. Dougherty[1]*, Sharona M. Atkins[1], Ana M. Franco-Watkins[2], Rick P. Thomas[3], Nicholas Lange[3] and Brandon Abbs[4]

[1] Decision Attention and Memory Lab, Department of Psychology, University of Maryland, College Park, MD, USA
[2] Evolution and Human Behavior, Department of Psychology, Auburn University, Auburn, AL, USA
[3] Auditory Neuroethology Lab, Department of Psychology, University of Oklahoma, Norman, OK, USA
[4] Department of Psychiatry, Harvard Medical School, Harvard University, Boston, MA, USA

Edited by:
Lael Schooler, Max-Planck Institute for Human Development, Germany

Reviewed by:
Ben Newell, University of New South Wales, Australia
Joseph G. Johnson, Miami University, USA
Lael Schooler, Max-Planck Institute for Human Development, Germany

***Correspondence:**
Michael R. Dougherty, Department of Psychology, University of Maryland, College Park, MD 20742, USA.
e-mail: mdougherty@psyc.umd.edu

We tested the predictions of HyGene (Thomas et al., 2008) that both divided attention at encoding and judgment should affect the degree to which participants' probability judgments violate the principle of additivity. In two experiments, we showed that divided attention during judgment leads to an increase in subadditivity, suggesting that the comparison process for probability judgments is capacity limited. Contrary to the predictions of HyGene, a third experiment revealed that divided attention during encoding leads to an increase in later probability judgment made under full attention. The effect of divided attention during encoding on judgment was completely mediated by the number of hypotheses participants generated, indicating that limitations in both encoding and recall can cascade into biases in judgments.

Keywords: working memory, probability judgment, hypothesis generation, support theory

IMPLICATIONS OF COGNITIVE LOAD FOR HYPOTHESIS GENERATION AND PROBABILITY JUDGMENT

A fundamental assumption of many models of judgment is that judgment is dependent on memory. This assumption was first realized by Tversky and Kahneman (1974) in their work on the availability and representativeness heuristics, as well as subsequent work on confidence and probability judgments (Tversky and Koehler, 1994; Dougherty et al., 1999; Dougherty, 2001; Sieck and Yates, 2001; Goldstein and Gigerenzer, 2002; Juslin and Persson, 2002), hypothesis generation (Gettys and Fisher, 1979; Weber et al., 1993; Dougherty and Hunter, 2003a,b; Dougherty and Sprenger, 2006; Thomas et al., 2008), and even choice (Weber et al., 2007).

Despite the importance of memory for judgment processes, a detailed analysis of the role of memory processes in judgment is lacking. Indeed, even some influential models of probability judgment fail to explicate the precise nature of the memory processes that underpin judgments. For example, in Tversky and Koehler's (1994) support theory, the perceived support for a particular hypothesis is assumed to be based on an underlying memory variable. However, rather than specifying these underlying memory variables, Tversky and Koehler (1994) suggested that support was assessed using judgmental heuristics such as availability and representativeness. Similarly, Goldstein and Gigerenzer (2002) proposed the recognition heuristic as a mechanism for choice, but did not specify the memorial basis of the recognition heuristic (Pleskac, 2007; Dougherty et al., 2008). On the one hand, it seems reasonable to study judgment processes independent of their memorial inputs, and much has been learned from such investigations. On the other hand, the predictions of any judgment model can only be as good as the assumptions upon which it rests. If one assumes that memory processes provide the input to the judgment process, then a complete understanding of the judgment process necessitates that we understand the pre-decisional processes involved in memory. Accordingly, appealing to vague heuristics such as availability and representativeness as *mechanisms of memory* do little to further our understanding of judgment.

The last several years has seen a growth in models of judgment grounded in memory theory. Dougherty et al. (1999) extended a model of recognition memory, Minerva 2, to deal with conditional probability judgment and categorization. Juslin and Persson (2002) extended a model of categorization to model probability judgment and confidence. Pleskac (2007) specified Goldstein and Gigerenzer's (2002) recognition heuristic within the context of signal detection theory, and Schooler and Hertwig (2005) did so within the context of ACT-R. Finally, Thomas et al. (2008) proposed an integrative model, HyGene, that links the processes of hypothesis generation, information search, and judgment to both long-term memory retrieval and working memory. Still, though, there have been relatively few experimental studies investigating the interrelationship between memory and judgment.

In this paper we bring together research from the areas of decision making, attention, and memory. Our view is that memory processes serve as input into the processes of judgment and decision making. Thus, errors and biases that crop up in the memory retrieval process, or constraints placed on decision makers that limit their time or cognitive resources, will cascade into errors and biases in judgment and decision making. Specifically, we were interested in examining the relationship between working memory, divided attention, and probability judgment in an inductive inference task that required participants to generate hypotheses from memory (Gettys and Fisher, 1979; Mehle, 1982; Dougherty and Hunter, 2003a,b). Such generation processes underlie a number of real-world judgment tasks, including medical diagnosis (Weber et al., 1993), accountants' generation of going concern problems (Libby, 1985), and fault generation by power-plant operators (Patrick et al., 1999). To our knowledge, no research has yet examined the effect of dividing attention on probability judgment.

We start with the assumption that eliciting a probability judgment of an elementary event prompts people to generate its logical alternatives from long-term memory. For example, if a clinician learns that a patient has shortness of breath and congestion, and is asked to assess the likelihood that the patient has bronchitis, we assume that the clinician will generate logical contenders to bronchitis (e.g., asthma and emphysema). These alternatives are assumed to form the evaluative basis for comparing the strength of evidence in favor of bronchitis to its contenders. Thus, we expect memory variables to influence the outcome of the judgment process to the extent that they affect which and how many hypotheses are retrieved from memory and included as part of the evaluation process.

The second assumption concerns the precise nature of the judgment process. Following Tversky and Koehler (1994), we assume that participants utilize a comparison process to derive their probability judgments. Formally, this comparison process is given by Eq. 1 (see Tversky and Koehler, 1994):

$$p(H_a, H_b) = \frac{s(H_a)}{s(H_a) + s(H_b)}, \tag{1}$$

where $s(H_a)$ and $s(H_b)$ correspond to the strength of evidence for hypotheses H_a and H_b respectively. Returning to the physician assessing the likelihood of bronchitis (H_a), the set of alternatives to bronchitis (H_b) consists of the implicit disjunction of not bronchitis. Tversky and Koehler (1994) found that people's judgments tend to be subadditive; the probability of an implicit disjunction tends to be lower than the sum of the probabilities assigned to its elements. For example, if one were to judge *p*(not bronchitis, bronchitis) it would be judged as less likely than the sum of the probabilities assigned to *p*(asthma, bronchitis), *p*(emphysema, bronchitis), and *p*(all other non-bronchitis possibilities). Thus the judged probability of the inclusive hypothesis, *p*(not bronchitis, bronchitis) is subadditive with respect to the sum of the judged probabilities of its elements.

We assume that merely being asked to judge the likelihood of bronchitis (H_a) prompts the decision maker to generate hypotheses implicit in the set of H_b to be included in the comparison process. Note that according to Eq. 1, the judged probability of H_a should be negatively related to the number and strength of the alternatives from H_b included in the comparison process. Considerable research suggests that the comparison process is well characterized in terms of Eq. 1 (Dougherty et al., 1999; Thomas et al., 2008; Tversky and Koehler, 1994; Windschitl and Wells, 1998).

Dougherty and colleagues have argued that hypothesis generation and working memory processes constrain how many and which hypotheses are included in the comparison process (Dougherty and Hunter, 2003a,b; Dougherty and Sprenger, 2006; Sprenger and Dougherty, 2006; Thomas et al., 2008). For example, Dougherty and Hunter (2003a), Dougherty et al. (1997) demonstrated that people tended to generate strong (high-probability) alternatives as opposed to weak (low-probability) alternatives. Moreover, the number of hypotheses included in the comparison process was positively related both to individual differences in working memory capacity and to the amount of time allowed for generation (Dougherty and Hunter, 2003b; Sprenger and Dougherty, 2006). In all these cases, the number and strength of alternatives generated and included in the comparison process was negatively correlated with the magnitude of participants' judgments.

More recently, Thomas et al. (2008) proposed the HyGene model to account for the relationship between long-term memory retrieval (i.e., hypothesis generation), working memory, and judgment. HyGene is able to account for a variety of effects in the probability judgment literature (see Dougherty et al., 2010), including the alternative outcomes effect (Windschitl and Wells, 1998), subadditivity effects (Tversky and Koehler, 1994), and the negative correlation between individual differences in working memory span and judgment (Dougherty and Hunter, 2003a,b). According to HyGene, the negative correlation between individual differences in working memory span and judgment arises from limitations on the number of alternatives included in the comparison process: participants with higher working memory capacities are assumed to include more alternatives in the comparison process. Within HyGene, the number of hypotheses included in the comparison process can be determined both by working memory constraints on the comparison process and by how well information is encoded in long-term memory. The capacity hypothesis postulates that working memory capacity constrains the *number* of hypotheses that one can include in the comparison process, irrespective of the number that are retrieved from long-term memory. Thus, it is theoretically possible for one to retrieve a relatively large number of hypotheses from long-term memory, yet include only a few of these hypotheses in the comparison process due to working memory limitations.

HyGene postulates that the number of alternatives included in the comparison process will be constrained by how many alternatives can actually be retrieved from long-term memory. This possibility suggests that probability judgments that require one to generate information from memory will be sensitive to how well that information was initially encoded in memory. Considerable research indicates that encoding is a resource-dependent process (Craik et al., 1996; Naveh-Benjamin et al., 1998; Fernandes and Moscovitch, 2000; Kane and Engle, 2000). For example, Craik et al. (1996) showed that divided attention during encoding (compared to full attention) led to substantial decreases in later retrieval, whereas divided attention during retrieval led to minimal decrements in the number of items retrieved. This finding prompted Craik et al. (1996) to suggest that the process of encoding was modulated by attention whereas retrieval was obligatory and protected from the effects of divided attention (for a different view, see Fernandes and Moscovitch, 2000).

The idea that attention is necessary for encoding raises the possibility that differences in judgment magnitude between participants with large and small working memory capacities arises from differences in the encoding process, rather than constraints on the comparison process. If the retrieval of the alternatives to the to-be-judged items is dependent on how well they were encoded in memory prior to the judgment task, then one would expect any sort of memory-dependent judgment to be sensitive to how well the alternatives were encoded[1]. Given the evidence that individuals with larger working memory capacity appear to have better encoding processes compared to individuals with smaller working memory capacity, any correlation between working memory capacity and memory, as well as working memory capacity and judgment, may be the result of encoding processes. For example, better encoding of alternatives for high-span participants would lead them to retrieve more alternative hypotheses (i.e., elements from H_b) for inclusion in the comparison process, which according to Eq. 1, could lead to lower judgments. The differential encoding explanation is consistent with research showing that participants with larger working memory capacity devote more attentional resources to encoding as a means of dealing with proactive interference (Kane and Engle, 2000). Taken together, differences in the number of hypotheses retrieved between high- and low-span participants, as well as differences in the magnitude of probability judgments, may result from encoding processes and not constraints on the number of hypotheses that one can include in the comparison process.

SIMULATING THE EFFECT OF WORKING MEMORY AND ENCODING ON PROBABILITY JUDGMENT

To illustrate the relationship between working memory, encoding, and probability judgment, we used HyGene[2] to simulate how manipulating working memory and encoding quality might affect the magnitude of people's probability judgment. HyGene is based on three core principles: (1) data observed in the environment serve as retrieval cues that prompt the retrieval of hypotheses from long-term memory. (2) The number of hypotheses that one can actively entertain is constrained by working memory capacity and task characteristics. (3) Hypotheses maintained in working memory are used as input into a comparison process to derive probability judgments. Principles 1 and 2 describe what we referred to as pre-decisional processes: they determine *what* and *how much* information is generated from memory and included in the judgment process. While these processes can be considered independent of the judgment process, the judgment process clearly depends on these pre-decisional processes. Specifically, retrieved hypotheses serve as the evaluative basis for determining judged probability. Model details are provided in the Section "Appendix."

To illustrate HyGene's predictions we manipulated the frequency of alternative hypotheses, working memory capacity, and encoding quality within HyGene. **Table 1** briefly describes HyGene's parameters and presents the parameter values used for the simulation. The two parameters relevant for the current simulations are working memory capacity (ϕ) and encoding level (L). Within HyGene, working memory capacity is defined as the number of hypotheses that can be maintained in an active state at any given time. The effect of working memory capacity on judgment was assessed across two levels to simulate low capacity ($\phi = 2$) and high capacity ($\phi = 4$). To simulate the effect of manipulating encoding, we varied the encoding parameter, L (the probability that a feature of an experienced event is encoded into the corresponding memory trace), across a range of five levels: $L = 0.5, 0.6, 0.7, 0.8, 0.9$. Prior work has illustrated that judgment magnitude is affected by the strength of the alternatives one considers (see Dougherty and Hunter, 2003a). Thus, we manipulated the strength of alternatives

[1] Throughout the paper, when we refer to "better encoding" or "higher encoding levels," we mean that individuals have stored more item and/or context information for items/alternatives that they are attending. In contrast, when we refer to "worse encoding," or "decreases in encoding," we mean that individuals have stored less item and/or context information for items/alternatives that they are attending.

[2] We used the HyGene model for simulations because it is the only model that we know of that has been designed to model the realistic situation in which people must retrieve hypotheses from long-term memory and feed these hypotheses into long-term memory.

Table 1 | HyGene parameters, definitions, boundary values, and values used in the simulation.

Parameter	Definition	Boundaries	Values used in simulation
L	Quality of memory encoding: the probability that each feature in the experienced event is encoded into the corresponding memory trace vector.	$0 \leq L \leq 1$	$L = 0.5$ $L = 0.6$ $L = 0.7$ $L = 0.8$ $L = 0.9$
TMAX	A retrieval parameter that determines how long the model searches semantic memory. Generation terminates when the total number of retrieval failures (the model fails to retrieve a novel hypothesis) exceeds TMAXTMAX can be used to model task characteristics such as time pressure, and individual variables such as effort or motivation.	TMAX ≥ 0	TMAX = 10
A_c	The activation criterion for memory traces in episodic memory. Episodic traces are placed in the activated subset if their activation exceeds this threshold, A_c.	$0 \leq A_c \leq 1$	$A_c = 0.166$
ϕ	This working memory capacity parameter specifies the upper limit of how many hypotheses can be held in working memory.	$\phi \geq 0$	$\phi = 2$ (low working memory capacity) $\phi = 4$ (high workingmemory capacity)
$\text{Act}_{\text{Min}H}$	The activation criterion for hypotheses to be placed in working memory. $\text{Act}_{\text{Min}H}$ is always set to 0 initially, and then is dynamically updated based on the activation values of hypotheses in the SOC.	$\text{Act}_{\text{Min}H} \geq 0$	$\text{Act}_{\text{Min}H} = 0$ (to start)

using two different frequency distributions to examine its effect on judgment. For the balanced distribution condition, eight hypotheses were stored in memory with trace frequencies of 10, 5, 5, 4, 4, 4, 4, and 1. For the unbalanced condition, eight hypotheses were stored in memory with trace frequencies of 15, 10, 7, 1, 1, 1, 1, and 1. These distributions closely approximate the form of those used in the behavioral experiments presented in this paper[3]. Each simulation consisted of 1,000 trials.

Figure 1 presents the simulation results for the effect of distribution on the sum of the eight probability judgments, as a function of working memory capacity. Note that to be additive, the model's eight judgments should sum to 100%, and the degree to which they exceed this demonstrates that the predicted judgments are subadditive. HyGene predicts that judgments will have higher absolute magnitude (and thus will produce greater subadditivity) in the balanced distribution than in the unbalanced distribution. Also, HyGene predicts a modest decrease in judgment magnitude and subadditivity as working memory capacity increases. These model predictions are consistent with the negative correlation between individual differences in working memory capacity and judgment magnitude found in prior work (Dougherty and Hunter, 2003a,b).

Figure 2 plots HyGene's probability judgment predictions as a function of encoding quality for both balanced and unbalanced distributions. HyGene makes two important predictions regarding the effect of encoding on hypotheses generation and judgment. First, HyGene predicts that participants will retrieve more alternatives to the focal hypothesis under conditions of high encoding (not shown). Thus, hypothesis generation is expected to show a positive relationship with encoding. Second, and less intuitively, HyGene predicts that high encoding will lead to *increases* in judgment magnitude and subadditivity. This later prediction is non-intuitive, since the formula for the comparison process (Eq. 1) predicts that increases in the number of hypotheses included in the denominator should lead to a decrease in the perceived probability of the focal hypothesis and decreased subadditivity. HyGene predicts increased judgment magnitude because increases in encoding quality lead to a disproportionate increase in the strength of the focal relative to the alternatives. Although the model generates more alternatives with increases in encoding quality, the increase in the strength of the focal hypothesis overwhelms the effect of including more alternatives in the comparison process. The non-intuitive nature of this prediction serves as a strong test of the version of HyGene presented by Thomas et al. (2008).

Previous research found that individual differences in working memory capacity are negatively correlated with probability judgment magnitude (e.g., Dougherty and Hunter, 2003a,b). One hypothesis, supported by the HyGene simulation, is that increased capacity is related to an increased number of alternatives in the comparison process, and consequently decreased judgments. However, another plausible explanation is that participants high in working memory capacity have better knowledge of extensional rules of probability (e.g., additivity within a sample space), mathematical ability, or general knowledge (individuals high in working memory capacity tend to perform better on measures of intelligence and general ability, including verbal and quantitative SATs; Rohde and Thompson, 2006). In the present experiments, we manipulated cognitive load, rather than exclusively using a correlational design. If high capacity participants make lower probability judgments because of quantitative knowledge, they should be unaffected by reduced attention during the task. In contrast, if high capacity participants make lower judgments because of generation processes, dividing attention should reduce generation ability and lead them to make higher judgments than when using full attention.

EXPERIMENT 1

In Experiment 1, we tested the influence of divided attention at encoding and divided attention during retrieval on probability judgment. We manipulated cognitive load by requiring participants to remember strings of 1–8 letters while simultaneously engaging in a probability judgment task. We anticipated that increases in cognitive load would decrease the number of alternative hypotheses included in the comparison process and thereby lead to increases in judged probability.

[3] The absolute frequencies used in the simulation are of less importance than the relative frequencies.

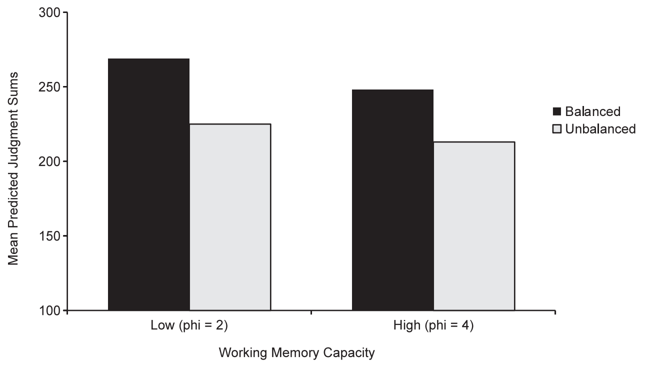

FIGURE 1 | Mean predicted sum of probability judgments as a function of the strength of alternative hypotheses and working memory capacity for the HyGene simulation. The distribution of alternatives for the balanced condition was: 10, 5, 5, 4, 4, 4, 4, 1. The distribution for the unbalanced condition was: 15, 10, 7, 1, 1, 1, 1, 1.

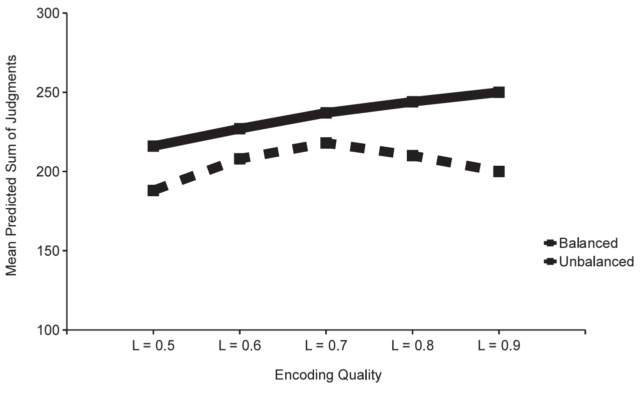

FIGURE 2 | Mean predicted sum of probability judgments as a function of encoding quality for balanced and unbalanced distributions for the HyGene simulation.

METHOD

Participants

Participants were 167 individuals who were recruited from the University of Maryland campus community by advertisements posted around campus and in the University newspaper. Participants were paid $10 for participating in the experiment.

Design

The experimental design was a randomized block design, with cognitive-load manipulated within participants across four levels. The cognitive-load task involved a "preload" manipulation that required participants to maintain 1, 4, 6, or 8 letters for short-term recall while simultaneously engaging in a probability judgment task.

Procedure

The entire experiment took place on computers and lasted 1.5 h. The experiment consisted of five main tasks: (1) a simulated restaurant task in which participants learned the frequency with which hypothetical customers ordered specific menu items at a diner, (2) a cognitive-load task in which participants were required to remember a list of 1, 4, 6, or 8 letters for short-term recall, (3) a judgment task, in which participants were asked to judge the likelihood that a given item would be ordered at the restaurant by a given person, (4) an implicit memory task which was used to test whether irrelevant items[4] were activated by the judgment task, and (5) the operation-span (o-span) task which was used to measure working memory capacity.

Prior to engaging in the experiment, participants completed a practice session that introduced them to the various tasks they would perform during the experiment. Participants practiced each of the individual tasks alone, and then practiced the tasks exactly as they would be performed in the experiment. The only difference between the experiment and the practice session was that the practice session used a simulated vacation scenario, in which participants learned the various travel destinations of hypothetical travelers, rather than the simulated restaurant scenario.

Restaurant task. For the simulated restaurant task, participants were told to pretend that they had a job waiting on tables at a particular country diner, and that they notice that four male customers come in regularly for four different meals and order from four different menus. There were no overlapping menu items across the four menus and each menu contained eight items. For example, Bob always ordered one of eight possible items from the breakfast menu, Steve always ordered one of eight possible items from the lunch menu, Tim always ordered one of eight possible items from the dinner menu, and Dan always ordered one of the eight possible items from the dessert menu. The restaurant task displayed each customer's order over the course of 50 consecutive days. The menu items were presented in random order with a frequency distribution of 10-10-8-8-4-4-3-3. The assignment of items to presentation frequency within a menu was counterbalanced such that each item was presented at each objective frequency an equal number of times across participants. The items learned in this phase served as the stimuli for the judgment task. The order of presentation of the meals (breakfast, lunch, dinner, and dessert) always occurred in the same order.

Following the simulated restaurant task, participants were told to imagine they had a new job waiting on tables in another restaurant, and they had four new customers who come in regularly and order from menus that are different from the first menus. These four new customers from the second diner were included to serve as "irrelevant" alternatives. This second simulated restaurant was identical to the first except for three features: (1) the four patrons at this diner were all female, whereas the patrons at the prior diner were all male, (2) none of the eight items on each of the four menus at the second diner were on the menus of the first diner, and (3) each of the eight items on the four menus was ordered exactly twice over the course of 16 simulated days. This second restaurant task was included for two reasons. First, it served as a distracter task between learning the to-be-judged items and the subsequent judgment task, thereby eliminating possible recency effects. Second, it served to provide a set of irrelevant stimuli that we hypothesized would interfere with the retrieval of the items from the main restaurant task. Ideally, participants should be able to discriminate and then inhibit items from the irrelevant menu when making their judgments. However, to the degree that participants are unable to discriminate and inhibit these irrelevant items, retrieval of these items might influence participants' judgments (Dougherty and Sprenger, 2006).

Cognitive-load task. Participants were required to retain a list of letters for later recall. The number of letters that had to be maintained for recall varied across four load levels: 1, 4, 6, or 8 letters. This manipulation of load constituted our main independent variable and was manipulated within participants.

Judgment task. For the judgment task, participants were required to make judgments about the likelihood that a given person would order a given item from the menu. Probability judgments were made by pressing 1 of 11 keys on the keyboard corresponding to probabilities ranging from 0 to 100% (in 10% increments). With one exception, the order of the presentation of the 32 menu items was randomized anew for each participant. The one exception to the judgment procedure outlined above concerns the first four items judged. All participants judged one frequency-10 item from each of the four menus to begin the task.

Implicit memory task. The implicit memory task consisted of 16 word fragments: 8 corresponded to menu items on the target menu and 8 to menu items from the irrelevant restaurant. Words from the menu were presented with approximately half the letters missing. The missing letters were randomly removed from each word. Participants were asked to complete the word fragments. The implicit memory task was included as a measure of what items (and how many) were primed by the judgment task.

The cognitive-load task was performed simultaneously with the judgment and implicit memory tasks. For each of the 32 items judged, participants first observed letters to remember for the cognitive-load task, then made a likelihood judgment, then recalled the letters for the cognitive-load task, and finally completed a word

[4] An "irrelevant" item was an item from a different menu that was not part of the set of possible hypotheses for the judgment at hand. For instance, if a participant was judging the likelihood that a diner would order eggs for breakfast, , and the participant thought of "French toast" as an alternative item, but French toast was not on this diner's menu, "French toast" would be defined as irrelevant.

fragment for the implicit memory task. As previously mentioned, cognitive load was manipulated within participants. Each participant made eight judgments under each possible load level (1, 4, 6, or 8 letters). Further, all items from one meal type were judged under the same load level. For example, a participant might judge each of the 8 breakfast items under a load of 4 letters each of the 8 dinner items under a load of 8 letters, each of the 8 dessert items under a load of 1 item, and each of the 8 lunch items under a load of 6 letters. The assignment of load level to menu (breakfast, lunch, dinner, and snack) was counterbalanced across participants.

Operation span. After completing the experimental task, each participant completed the operation-span (o-span) task as a measure of working memory span (Turner and Engle, 1989). The o-span task required that participants retain a growing list of words while verifying arithmetic problem solutions. For example, on successive presentations participants would be shown: "$(4 \times 3) - 3 = 9$? Door; $(4/2) + 3 = 7$? Shoe." Participants were required to read the arithmetic problem and its solution aloud, verify whether the solution was true, and then read the word aloud. After saying the word, the experimenter advanced to the next operation-word pair. This continued until the participant was prompted to recall the words from that set in the order in which they were presented. Participants were presented with 15 sets of equation-word pairs with set sizes ranging from 2 to 6. Each set size occurred three times randomly throughout the task. Performance on the o-span task was computed by the number of words recalled in the correct serial position for which the corresponding arithmetic problem was correctly verified. The maximum possible score was 60, with higher scores representing larger working memory capacity. A detailed description of the operation-span task is presented in Turner and Engle (1989).

RESULTS AND DISCUSSION

Our primary analyses center on the effect of cognitive load on judged probability. However, several other aspects of the data are worth noting. **Table 2** presents the mean number of letters recalled for the cognitive-load task, number of relevant and irrelevant alternatives identified in the implicit memory task, judgment sums, gamma correlations, and the o-span scores.

Cognitive-load task

As a manipulation check, we examined the percent correct recall on the cognitive-load task. There were three main findings: first, the percent of letters correctly recalled on the cognitive-load task decreased significantly as the load level increased, $F(3,163) = 284.52$, $p < 0.05$. This indicates that the cognitive-load manipulation was successful. Second, working memory span was a significant predictor of the percent of correct letters recalled on the cognitive-load task, $F(1,165) = 37.15$, $p < 0.05$. Finally, the effect of cognitive-load interacted with working memory span, $F(3,163) = 11.63$, $p < 0.05$, in that participants with higher working memory span were less affected by the increase in cognitive load.

Implicit memory task

The mean number of set relevant and set irrelevant items identified in the implicit memory task for each level of cognitive load is shown in **Table 2**. Analysis of covariance using o-span as the covariate revealed a main effect of relevance on implicit memory performance, $F(1,162) = 15.42$, $p < 0.05$. Participants correctly identified more word fragments belonging to the relevant set than to the irrelevant set. In addition, working memory span was a significant positive predictor of implicit memory performance, $F(1,162) = 6.02$, $p < 0.05$. However, there was no effect of cognitive load on implicit memory performance, and no significant interactions between cognitive load, working memory span, and relevance (all p's > 0.20). For this reason, all subsequent covariance analyses used the mean number of irrelevant alternatives identified collapsed across cognitive-load level.

Probability judgments

We analyzed two aspects of judgment accuracy: (1) the degree to which participants' judgments were subadditive (summed to greater than 100%), and (2) the degree to which participants' judgments discriminated among different levels of objective probability (i.e., judgments' relative accuracy). These two measures capture different components of judgment accuracy. Subadditivity captures the degree to which participants' judgments satisfy the additivity axiom of probability theory. However, judgments can be additive without accurately discriminating between events with different objective probabilities (see Dougherty and Sprenger, 2006).

Table 2 | The mean number of relevant and irrelevant alternatives identified in the implicit memory task as a function of cognitive-load level, the percentage of letters recalled in the cognitive-load task as a function of cognitive-load level, mean judgment sums as a function of cognitive-load level, mean gamma correlations as a function of cognitive-load level, and mean operation-span scores for Experiment 1.

	Cognitive load level				
	1 Letter	**4 Letters**	**6 Letters**	**8 Letters**	**o-Span**
Percent of letters recalled on cognitive-load task	0.91 (0.02)	0.96 (0.01)	0.84 (0.02)	0.67 (0.02)	
IRR	3.46 (0.13)	3.51 (0.14)	3.46 (0.13)	3.40 (0.14)	
REL	4.33 (0.14)	4.29 (0.14)	4.40 (0.13)	4.23 (0.13)	
Judgment sums	261.8 (8.3)	265.4 (9.2)	264.7 (8.4)	271.5 (9.6)	
Gamma	0.31 (0.03)	0.32 (0.03)	0.23 (0.03)	0.28 (0.03)	
o-Span					27.86 (0.72)

SE are presented in parentheses. o-Span, operation-span; IRR, number of irrelevant alternatives identified in implicit memory word-fragment task, REL, number of relevant alternatives identified in the implicit memory word-fragment task; Sums, sum of probability judgments.

Thus, we also examined the relative accuracy of each participant's judgments by computing a gamma correlation between each participant's subjective probability judgments and the corresponding objective probability values.

Table 2 presents the mean judgment sums for the four load conditions. As can be seen, participants' judgments were well above 100%, indicating that participants' judgments were subadditive. This finding is consistent with a growing body of research that participants often give judgments that sum to more than the probability of the implicit disjunction (Tversky and Koehler, 1994; Dougherty and Hunter, 2003a,b; Dougherty and Sprenger, 2006; Sprenger and Dougherty, 2006).

Our main experimental hypothesis was that the magnitude of participants' judgments should be related both to individual differences in working memory span and to cognitive load. Specifically, we expected that the magnitude of participants' judgments would be negatively correlated with working memory span and would increase as cognitive-load increased. **Table 3** presents the Pearson correlation coefficients between the sum of participants' judgments at each level of cognitive load and three predictor variables: o-span, mean number of irrelevant alternatives identified in the word-fragment task, and the mean number of relevant alternatives identified in the word-fragment completion task. Consistent with our hypotheses and prior research, working memory span was negatively correlated with the sum of participants' probability judgments. Additionally, judged probability was affected by cognitive load; inspection of the mean sums in **Table 2** reveals that there was about a 10% increase in the magnitude of participants' judgments between the load 1 and load 8 conditions. Analysis of covariance using o-span score and the mean number of irrelevant alternatives identified in the word-fragment task as covariates, revealed a significant effect of cognitive load on judgment, $F(3,161) = 3.13$, $p < 0.05$, with o-span as a significant predictor of judgment, $F(1,163) = 8.32$, $p < 0.05$. The number of irrelevant alternatives identified in the word-fragment task was not a significant predictor, $F(1,163) = 0.04$, $p > 0.20$[5]. This pattern of results indicates that increases in cognitive load lead to an increase in judgment magnitude, and the presence of irrelevant alternatives did not affect judgment magnitude.

Dougherty and Sprenger (2006) argued that decreases in working memory capacity should lead to increases in judgment magnitude without having a concomitant effect on relative accuracy. We measured relative accuracy by computing a gamma correlation between participants' subjective probability judgments and their corresponding objective probability values. Mean gamma correlations are presented in **Table 2**. Analysis of covariance using o-span and the number of irrelevant alternatives identified in the word-fragment task as covariates revealed no effect of cognitive load on relative accuracy, and no significant relationships between o-span and gamma or between the number of irrelevant alternatives identified and gamma (all *p*'s > 0.20). The results suggest that changes in judgment magnitude do not necessarily lead to changes in the rank order correlation between objective and subjective probabilities.

EXPERIMENT 2

Experiment 1 revealed a main effect of cognitive load on the sum of participants' judgments. However, the effect of cognitive load on judgment was relatively small. One possibility for this small effect was our use of the cognitive-load manipulation, rather than a concurrent processing task (Rosen and Engle, 1997). Thus, we chose to replicate Experiment 1 using a concurrent dual task during judgment that would place a greater attentional demand on participants. In addition to replicating the effect of cognitive load on probability judgments, Experiment 2 was designed with two additional goals. First, we manipulated whether participants were primed with a relevant or an irrelevant alternative, rather than measuring the number of irrelevant alternatives generated by individuals *post hoc*. Second, we manipulated whether the distribution of irrelevant alternatives consisted of "strong" or "weak" items, where strength was defined by the items' absolute frequency.

METHOD

Participants

Participants were 73 undergraduate students enrolled in psychology courses at the University of Maryland. Participants received course extra-credit for participating in the experiment.

Design

The experimental design was a 2 (cognitive load: high versus low) × 2 (prime type: relevant versus irrelevant) × 2 (strength of irrelevant alternatives: strong versus weak) mixed factorial, with cognitive load and prime-type manipulated within participants and strength of irrelevant alternatives manipulated between participants.

Procedure

The general procedure and instructions for Experiment 2 was the same as Experiment 1, with the following exceptions. First, after learning the four relevant distributions, participants learned an

Table 3 | Pearson *r* correlations between operation-span, the number of relevant and irrelevant alternatives identified in the implicit memory task, and judgment sums as a function of cognitive-load level for Experiment 1.

	Sum 1 letter	Sum 4 letter	Sum 6 letter	Sum 8 letter	IRR	REL
o-Span	–0.284**	–0.238**	–0.187*	–0.218**	0.163*	0.184*
IRR	–0.081	–0.070	0.016	–0.056		
REL	–0.042	–0.004	0.001	–0.013		

*o-Span, operation-span; IRR, number of irrelevant alternatives identified in the implicit memory word-fragment task; REL, number of relevant alternatives in the implicit memory word-fragment task; Sum, sum of probability judgments. *p < 0.05, **p < 0.01.*

[5] IRR was uncorrelated with judgment, but was correlated with o-span. Although not a significant predictor of judgment sums itself, IRR served as a suppressor variable on o-span, and therefore was necessary for inclusion in the model.

irrelevant distribution that consisted of either eight weak items (each item presented two times during study) or two strong items (each item presented eight times during study). The asymmetry in the number of irrelevant alternatives allowed us to manipulate strength without varying the amount of time separating when participants studied the relevant distribution and when making their judgments.

Second, rather than engaging in a cognitive load task participants in Experiment 2 performed a concurrent finger-tapping task. Participants placed the four fingers of their right hand on the "j", "k", "l", and ";" keys in typing position (i.e., index finger on the "j" key, middle finger on the "k" key, ring finger on the "l" key, and littlest finger on the ";" key). Participants were required to press each key when they heard a tone associated with that key. Each key was associated with a different-pitched tone. The lowest pitch was associated with the "j" key, and the highest pitch was associated with the ";" key. In the low cognitive-load condition, the tone sequence always began with the lowest-pitched tone and incremented sequentially to the highest-pitched tone. Thus, participants pressed keys in order from index finger, to middle finger, to ring finger, and finally to littlest finger. The sequence then began again with the lowest-pitched tone and continued in the same cycle. In the high cognitive-load condition the tone sequence was random. Thus, participants were required to pay more attention to the tones because there was no predictable pattern to the sequence of tones. Participants had 500 ms to respond before the next tone played. If participants did not respond in time, the trial was counted as incorrect and the next trial began. The pairing of auditory tones with finger presses departs somewhat from prior uses of the figure-tapping task. For example, Kane and Engle (2000) manipulated whether participants tapped fingers sequentially in order (index, middle, ring, pinky) or in an alternating sequence (e.g., index, ring, middle, littlest). Although the alternating sequence is clearly more difficult than the sequential sequence, it is also susceptible to practice effects. Thus, the degree to which the alternating sequence actually divides ones attention will actually decrease throughout the experiment. In contrast, our random version of the finger-tapping task is less susceptible to practice effects, because participants cannot anticipate which finger will need to be pressed on any given trial.

The third main difference between Experiments 1 and 2 pertained to the timing of the implicit memory task. In Experiment 1, participants completed a word-fragment task *after* making each probability judgment. In Experiment 2, participants completed the word-fragment task *before* making each probability judgment. Thus, the word-fragment completion task in Experiment 2 served to prime participants to retrieve either items from the relevant distribution or items from the irrelevant distribution. The relevant primes were the two frequency-3 items from the relevant menu list. The irrelevant primes were two of the items from the irrelevant menu list. Half of the judgments were made following relevant primes and half were made following irrelevant primes. Judgments were always primed with an item from the meal being judged (i.e., if participants were asked the probability that Bob would order pancakes, the prime would be from the breakfast menu).

The final difference between Experiment 1 and 2 was that judgments were made verbally, rather than by responding with a key press. Participants provided a numerical probability judgment between 0 and 100 as in Experiment 1, but did so by stating the number aloud. This was done primarily to reduce motor-response interference since the cognitive-load task required a motor response. An experimenter sat with the participant during the judgment phase and coded the verbal probability judgments.

In the experiment, two distributions were judged under high cognitive load and two distributions were judged under low cognitive load. The level of cognitive load was blocked and counterbalanced across participants such that half of the participants judged two distributions under high cognitive load first followed by two distributions under low cognitive load, whereas the remaining participants completed the experiment in the reverse order.

Practice session

As with Experiment 1, all participants engaged in a practice session to familiarize themselves with the individual tasks and to practice the tasks in the sequence they occurred in the main experiment.

RESULTS

Manipulation check: finger-tapping task

There was a significant difference in finger-tapping accuracy (measured as percent correct) between the high and low cognitive-load conditions, $F(1,70) = 7.56$, $p < 0.05$, indicating that participants found the high-load condition more attentionally demanding. In addition, there was a significant effect of the strength of the irrelevant alternatives on finger-tapping accuracy, with participants in the "high irrelevant strength" condition performing significantly less well than participants in the "low irrelevant strength" condition, $F(1, 70) = 17.54$, $p < 0.0001$. As is clear from the finger-tapping data presented in **Figure 3**, there was also an interaction between cognitive load and strength of the irrelevant alternatives, $F(1,70) = 4.53$, $p < 0.05$. Univariate tests revealed that there was a main effect of strength of irrelevant alternative on finger tapping for participants in the high cognitive-load condition $F(1,71) = 11.61$, $p < 0.05$, but not for the low cognitive-load condition $F(1,71) = 3.05$, $p = 0.09$. These results offer the possibility that participants traded off processing resources dedicated to the finger-tapping task in order to effectively inhibit the irrelevant alternatives, particularly when the

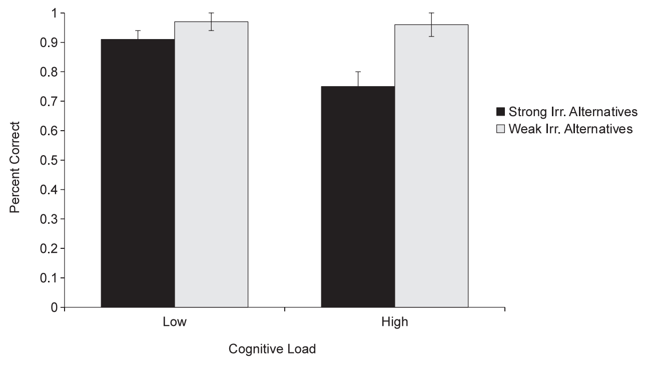

FIGURE 3 | Mean percent of correct finger-tapping responses as a function of cognitive-load level in the divided attention task for Experiment 2.

irrelevant alternatives were strong[6]. As in Experiment 1, working memory span was a significant predictor of performance on the cognitive-load task, $F(1,70) = 5.47$, $p < 0.01$, and working memory span interacted with cognitive load, $F(1,70) = 5.02$, $p < 0.05$, in that participants with higher working memory span were less affected by the increase in cognitive load.

Probability judgments

We manipulated whether participants were primed with an item from the relevant distribution or from the irrelevant distribution. The primes from the relevant distribution were those items that occurred with frequency-3. Note that priming participants with an item from the relevant distribution might strengthen those items, such that they might be rated as more likely after they have been used as a prime. Indeed, a preliminary analysis comparing the judged likelihood of items used as primes with those that were not used as primes indicated that participants gave higher absolute judgments for primed items than for non-primed frequency-3 items, $F(1,70) = 4.88$, $p < 0.05$. We therefore chose to exclude all of the frequency-3 items from our analyses to avoid confounding the effect of strengthening the two frequency-3 items with the effect of divided attention on judgment.

Our primary experimental hypothesis was that participants' judgments would be significantly higher under the high cognitive-load condition than in the low cognitive-load condition. Consistent with this hypothesis, there was a main effect of cognitive load on judgment sums, $F(1,69) = 5.56$, $p < 0.05$. Participants made higher absolute judgments in the high cognitive-load condition than in the low cognitive-load condition (as given by the marginal means in **Table 4**). Although cognitive-load affected the magnitude of participants' judgments, there was no effect of prime type on judgment, nor did any of the interactions reach significance (all p's > 0.15).

Table 4 | Mean judgment sums as a function of cognitive-load level, relevancy of prime, and strength of irrelevant alternatives for Experiment 2.

		Level of cognitive load task	
		Low	**High**
Irrelevant prime	Weak irrelevant alternatives	178.16 (14.66)	196.59 (16.46)
	Strong irrelevant alternatives	166.08 (14.66)	190.26 (16.46)
Relevant prime	Weak irrelevant alternatives	196.75 (14.43)	195.08 (15.54)
	Strong irrelevant alternatives	164.37 (14.84)	185.74 (15.54)
	Marginal means	177.10 (9.73)	192.02 (10.66)

SE are presented in parentheses. Sums, sum of probability judgments.

The third variable of interest in Experiment 2 was the strength of the irrelevant alternatives. Overall there was no effect of the strength of the irrelevant alternatives on judgment, nor did the strength of the irrelevant alternatives interact with prime type or cognitive load (all p's > 0.15)[7].

As with Experiment 1, we examined relative accuracy of participants' judgments by computing a gamma correlation for each participant between their subjective probability judgments and the corresponding objective probability values (excluding the frequency-3 items). **Figure 4** presents the mean of these gamma correlations. There was a main effect of prime type on relative accuracy, $F(1,68) = 4.63$, $p < 0.05$, and a marginally significant cognitive load by prime-type interaction, $F(1,68) = 3.33$, $p = 0.07$. The effect of prime type was limited to the high cognitive-load condition. However, counter to what one might expect relative accuracy was actually poorer when participants were primed with a *relevant* alternative than when they were primed with an *irrelevant* alternative. This result was surprising. We offer a possible interpretation of this finding in the Section "General Discussion."

EXPERIMENT 3

Experiments 1 and 2 offer evidence that divided attention at judgment perturbs the judgment process, and suggest that working memory capacity constrains the number of alternatives included in the comparison process. However, also important for this comparison process is the actual retrieval of alternatives from long-term memory. While retrieval has often been characterized as obligatory, the process of encoding is often assumed to be resource dependent (Craik et al., 1996; Naveh-Benjamin et al., 1998; Fernandes and Moscovitch, 2000; Kane and Engle, 2000). Indeed, considerable research has illustrated that divided attention at encoding leads to substantial decrements in later retrieval. Moreover, individual differences in working memory capacity apparently can impact encoding processes, as well as retrieval processes (Rosen and Engle,

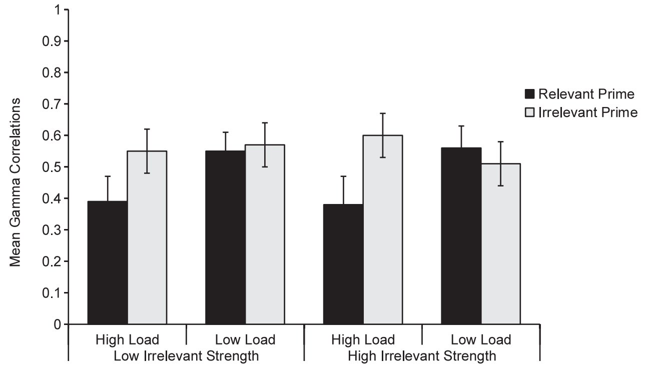

FIGURE 4 | Relative accuracy: mean gamma correlations between participants' subjective probability judgments and the corresponding objective probability values for Experiment 2.

[6] One possible way to probe for the proposed tradeoff is to examine the correlation between performance on the judgment task and performance on the finger-tapping task. Evidence for such a trade off would be given by showing that finger-tapping accuracy decreases the effect of irrelevant alternatives decreased. Analyses examining this correlation failed to find this relationship.

[7] Note that it is possible that the strength of irrelevant alternatives had no effects on judgment because we confounded strength with the number of irrelevant alternatives. The "strong irrelevant alternative" condition had only two alternatives, whereas the weak condition had eight alternatives.

1997; Kane and Engle, 2000). Thus, low working memory capacity can lead to poor encoding in long-term memory, which can cascade into poor retrieval (irrespective of whether one has full or divided attention at retrieval). If this assertion is correct, then encoding of judgment-relevant information should affect the number of alternatives participants can retrieve from long-term memory. Thus, we hypothesized that divided attention at encoding would lead to a decrease in the number of alternatives retrieved and used in the comparison process and that this decrease in retrieval would lead to an increase in judgment magnitude.

The basic experimental design was similar to that of Experiment 2. The primary differences were that (1) we manipulated the distribution of alternatives across two levels, (2) attention was divided at encoding instead of at judgment, (3) we required participants to engage in a recall task for two of the four menus, and (4) we administered an automated version of the o-span task (Unsworth et al., 2005).

METHOD

Participants

University of Maryland students (n = 101) completed the experiment in exchange for extra-credit in a psychology course.

Design

The experimental design was a 2 (cognitive load at encoding: high versus low) × 2 (distribution: balanced versus unbalanced) × 2 (recall versus no recall) mixed factorial, with divided attention at encoding manipulated between subjects and distribution type and recall manipulated within subjects.

Procedure

The general procedure in Experiment 3 was the same as Experiment 2 with the following exceptions. First, we manipulated the distribution of the alternatives. For two of the menus, the eight items were ordered with frequencies of: 15-10-7-1-1-1-1-1 (the *unbalanced* distribution). For the other two menus, the eight items were ordered with frequencies of: 10-5-5-4-4-4-4-1 (the *balanced* distribution). Second, we manipulated the recall of the menus. Each participant was asked to recall as many items as possible for one menu from the balanced distribution and for one menu from the unbalanced distribution. The recall of each particular menu was fully counterbalanced and the order of recall was randomly assigned. The recall task took place after the learning phase, and prior to the judgment phase. Third, divided attention was manipulated during the learning phase (rather than at judgment). The divided attention task was he same finger-tapping task as in Experiment 2.

The experiment consisted of five phases carried out in the following order: (1) practice (2) learning (this time while engaging in the finger-tapping task), (3) distractor task (to clear short-term memory: 30 s of finger tapping), (4) recall (completed under full attention), and (5) probability judgment (completed under full attention).

RESULTS

Manipulation check: finger-tapping task

As in Experiment 2, there was a significant difference in finger-tapping accuracy (measured as percent correct) between the high and low cognitive-load conditions, $F(1,95) = 10.91$, $p < 0.01$ indicating that participants found the high-load condition more attentionally demanding. Further, as in Experiments 1 and 2, working memory span was a significant predictor (marginally in this experiment) of performance on the cognitive-load task, $F(1,95) = 3.66$, $p = 0.059$.

Recall

We expected that participants would retrieve fewer alternatives in the high cognitive-load condition than in the low cognitive-load condition. **Figure 5** presents the mean number of alternatives recalled for the high and low cognitive-load conditions as a function of distribution type. Consistent with our hypothesis, there was a main effect of cognitive load on the number of alternatives recalled, $F(1,97) = 35.07$, $p < 0.01$. This finding is consistent with prior work in the memory literature, and indicates that our manipulation of divided attention at encoding successfully affected later recall.

We further examined the mean presentation frequency of the alternatives participants retrieved as a function of serial recall position. We hypothesized that participants would tend to recall the strongest (most frequently occurring) items earlier and subsequently recall weaker items. **Figure 6** shows that indeed, participants

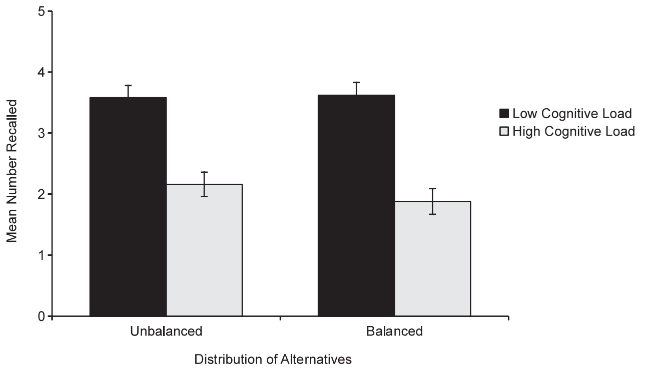

FIGURE 5 | Mean number of alternatives recalled for the high and low cognitive-load conditions as a function of distribution type for Experiment 3.

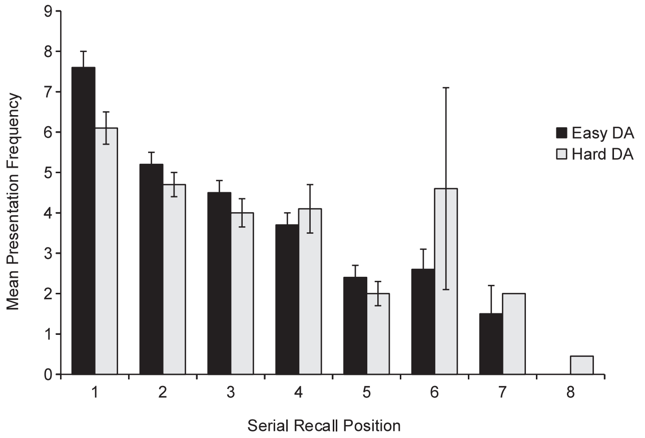

FIGURE 6 | Mean presentation frequency of recalled items as a function of serial recall position for Experiment 3.

tended to recall high frequency items first, and that the mean frequency of the recalled items decreased at each subsequent serial recall position.

Overall, the mean number of intrusions was low (High Cognitive Load: M = 0.69, SD = 1.56; Low Cognitive Load: M = 0.24, SD = 0.59), as was the number of repetitions (High Cognitive Load: M = 0.04, SD = 0.20; Low Cognitive Load: M = 0.04, SD = 0.20), and additions (High Cognitive Load: M = 0.31, SD = 0.65; Low Cognitive Load: M = 0.16, SD = 0.42).

Probability judgments

Figure 7 plots the mean sum of probability judgments for the high and low cognitive-load conditions for all four distributions. As expected, there was a main effect of cognitive load on mean sum of judgments, $F(1,97) = 6.55$, $p < 0.05$: High cognitive load during encoding led to increases in judgment sums even though participants made judgments under full attention. Taken together, both sets of results support the hypothesis that dividing attention during encoding leads to a decrease in the number of alternatives recalled and included in the comparison process, which consequently leads to increased probability judgments. However, note that this effect was not anticipated by HyGene, which predicted a lower judgment magnitude at lower levels of encoding.

We also expected that participants who recalled more alternatives for a given distribution would tend to make lower judgments, and have lower subadditivity for those judgments. Indeed, the sum of judgments was negatively correlated with the number of items participants recalled for that distribution of judgments [Balanced: $r(101) = -0.47$, $p < 0.0001$; Unbalanced: $r(101) = -0.33$, $p = 0.001$], and this was true across both high and low cognitive-load conditions. Further, we examined whether probability judgments were related to the strength (presentation frequency) of recalled items by summing the presentation frequencies of each item recalled for each distribution. Indeed, the sum of recall strength was negatively correlated with the sum of participants' probability judgments [Balanced: $r(101) = -0.37$, $p < 0.0001$; Unbalanced: $r(101) = -0.35$, $p < 0.0001$], and this was true across both high and low cognitive-load conditions. However, the strength of recalled items did not account for additional variance in probability judgments above and beyond the number of items participants recalled, balanced distribution: $F(1,97) = 0.32$, $p > 0.10$; unbalanced distribution: $F(1,97) = 2.31$, $p > 0.10$. An important question concerns whether recall mediates the relationship between divided attention at encoding and judgment. We tested this possibility by examining the effect of divided attention at encoding on judgment while controlling for the number of alternatives generated in the recall task. Consistent with the mediation account, the effect of divided attention at encoding on judgment was eliminated once the number of alternatives recalled was controlled both for the balanced condition, $F(1,97) = 0.00$, $p > 0.05$ and for the unbalanced condition, $F(1,97) = 0.04$, $p > 0.05$. This pattern of data suggests that probability judgments are dependent on the number of alternatives recalled from long-term memory and included in the judgment process.

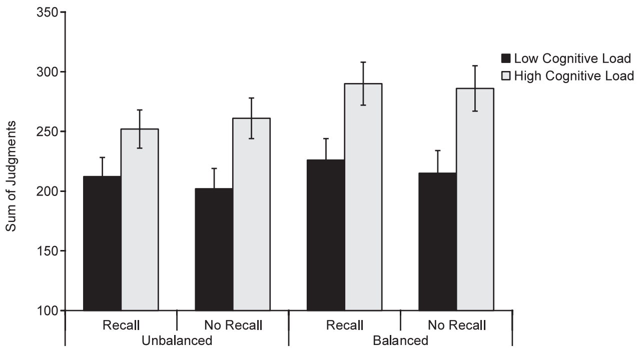

FIGURE 7 | Mean sum of probability judgments for the high and low cognitive-load conditions as a function of distribution type and recall condition for Experiment 3.

Table 5 presents correlations between working memory span, the number of alternatives recalled for the balanced and unbalanced conditions, and the sum of judgments for all distributions. Differences in working memory capacity correlated with both number of items recalled and the sum of participants' probability judgments. We further examined whether the effect of dividing attention at encoding on subadditivity still held when controlling for working memory span, and found that the effect was statistically significant both before and after controlling for individual differences in working memory span, $F(1,96) = 5.11$, $p < 0.05$. Working memory span was itself a significant predictor of judgment sums, $F(1,96) = 5.16$, $p < 0.05$. Thus, individual differences in working memory span do not appear to mediate the effect of dividing attention on judgment.

GENERAL DISCUSSION

Our thesis was that memory retrieval processes provide the input to the processes involved in judgment and decision making. To this end, we evaluated the impact of divided attention at both encoding and retrieval to examine how people make judgments of probability when the task requires that they generate hypotheses from long-term memory. To our knowledge, no research has yet examined the effect of dividing attention on probability judgment. It was hypothesized that when participants generate hypotheses from memory, working memory capacity would constrain the number of hypotheses they considered while making probability judgments. Additionally, we hypothesized that dividing attention at encoding would decrease the number of alternatives actually generated from long-term memory. In general, considering fewer

Table 5 | Pearson *r* correlations among automated operation-span, number of items recalled, and subadditivity for Experiment 3.

	Ao-span	Recall unbalanced	Recall balanced
JUDGMENT SUMS			
Unbalanced recall	–0.238*	–0.345**	–0.396**
Unbalanced no recall	–0.251*	–0.359**	–0.436**
Balanced recall	–0.229*	–0.408**	–0.466**
Balanced no recall	–0.230*	–0.358**	–0.476**

*Sums, sum of probability judgments; Ao-span, automated operation-span; Recall, number of items recalled. *p < 0.05, **p < 0.01.*

alternatives to a focal hypothesis increases one's perceived confidence in that hypothesis. Experiments 1 and 2 demonstrated that judgment magnitude increases when participants attention is reduced. This result extends previous research that found a negative relationship between working memory capacity and the magnitude of probability judgments (Dougherty and Hunter, 2003a,b; Sprenger and Dougherty, 2006). Our manipulation of cognitive load allow us to rule out the possibility that the prior correlations between judgment and working memory were due to idiographic differences in math ability, knowledge of probability theory, or encoding processes. Instead, our data indicate that dividing attention during the judgment process constrains the number of alternative hypotheses that one includes in the comparison process, leading to increases in judged probability. Moreover, these data are consistent with HyGene's prediction that lower levels of working memory capacity lead to increased probability judgments.

Dividing attention at encoding also affects subsequent judgments, leading both to a decrease in people's ability to retrieve alternatives and to an increase in judgment magnitude. Thus, errors and biases in the initial storage of information in long-term memory can cascade into errors and biases in judged probability. Such a dependence of judgment on memory is consistent with recent theoretical accounts that ground judgment processes in the underlying theoretical principles of memory (Dougherty et al., 1999; Thomas et al., 2008). Nevertheless, the effect of encoding on subadditivity was inconsistent with the HyGene simulations, which anticipated that increased encoding would lead to increased subadditivity.

Why does HyGene predict an increase in subadditivity with increased encoding in the face of increased retrieval? We believe that the effect stems from an overweighting of the focal hypothesis by the model: increases in encoding quality lead to a disproportionate increase in the strength of the focal relative to the alternatives. Although the model generates more alternatives with increases in encoding quality, the increase in strength resulting from increased encoding overwhelms the effect of including more alternatives in the comparison process. The net result, therefore, is that subadditivity is predicted to increase as a function increased encoding, even though the model also predicts increases in the number of alternative hypotheses generated. Addressing this misprediction within the vector based model proposed by Thomas et al. (2008) has proved challenging (see Dougherty et al., 2010 for further discussion on this point), though much more empirical and theoretical work is ongoing.

Taken together, the data presented here suggest a dual-role of working memory on judgment. On the one hand, working memory is important for judgment in that it constrains the number of alternatives one can include in the comparison process. This assertion is supported both by individual difference studies showing the negative correlation between judgment and working memory span, as well as by the present research, which shows that concurrent processing demands during the judgment task can lead to increases in judgment. On the other hand, working memory can also have an impact on judgment by affecting how well information is stored in long-term memory. Inasmuch as the input to the judgment process depends on the retrievability of information from long-term memory, factors that affect this retrievability will affect judgment. This assertion is clearly supported by the present research: divided attention during encoding led to a substantial increase in judgment magnitude, and this effect is completely mediated by people's ability to retrieve the alternatives to the to-be-judged event.

CONCLUSION

In many ways, our research parallels the debate regarding the effect of divided attention at retrieval. On the one hand, Craik et al. (1996; Naveh-Benjamin et al., 1998) suggested that retrieval processes are protected from the effects of divided attention. On the other hand, some researchers have suggested that attentional resources are needed for effective retrieval, particularly when the retrieval task requires the inhibition of irrelevant information (Rosen and Engle, 1997; Kane and Engle, 2000), or competition for the underlying memory representations by the secondary task (Fernandes and Moscovitch, 2000). While our research cannot definitively address whether divided attention during judgment impeded retrieval or placed restrictions on the number of hypotheses included in the comparison process, it clearly showed that judgment is resource demanding when the task entails a comparison of hypotheses generated from memory. This is not to say that judgment processes are always resource demanding. For example, it is quite likely that simpler judgment tasks, such as the statement verification task (Wallsten and González-Vallejo, 1994; Wallsten et al., 1999) require little cognitive resources (Pleskac et al., 2008). At the same time, our research also concurs with findings showing that divided attention at encoding leads to decrements at retrieval (Craik et al., 1996; Naveh-Benjamin et al., 1998). However, the implications for divided attention at encoding go beyond simple memory tasks and affect the higher-level judgment and decision processes that work on the output of the memory system (cf. Pleskac et al., 2008).

In sum the present research is consistent with the view that capacity limitations constrain the number of alternative hypotheses included in a support-theory like comparison process. Indeed, considerable research has revealed that people's probability judgments are subadditive across a range of tasks, including confidence memory for personal events (Mulford and Dawes, 1999), perceived probability of causes of death (Tversky and Koehler, 1994), and assessments of gambles (Fox and Tversky, 1998). The present research suggests that estimates of probabilities in these contexts will be inflated when cognitive capacity is low. Importantly, cognitive capacity can be low due to idiographic differences amongst decision makers, or by placing the decision maker under cognitive load. Thus, the present work adds to a growing body of literature illustrating the importance of working memory processes in hypothesis generation and probability judgment: both lower working memory capacity and task-induced increases in cognitive load lead to increases in the perceived probability of events. We argue that cognitive-load reduces the number of alternatives used in the comparison process, which in turn leads to an increase in the judged likelihood of the focal hypothesis. In this way, our research suggests that errors, biases, or limitations in the memory retrieval process can cascade into errors and biases in judgment.

REFERENCES

Craik, F. I., Govoni, R., Naveh-Benjamin, M., and Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *J. Exp. Psychol. Gen.* 125, 159–180.

Dougherty, M. R., Franco-Watkins, A. M., and Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychol. Rev.* 115, 199–211.

Dougherty, M. R., and Sprenger, A. (2006). The influence of improper sets of information on judgment: how irrelevant information can bias judged probability. *J. Exp. Psychol. Gen.* 135, 262–281.

Dougherty, M. R., Thomas, R. P., and Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. *Psychol. Learn. Motiv.* 52, 300–342.

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *J. Exp. Psychol. Gen.* 130, 579–599.

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

Dougherty, M. R. P., Gettys, C. F., and Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organ. Behav. Hum. Decis. Process* 70, 135–148.

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

Dougherty, M. R. P., and Hunter, J. E. (2003b). Probability judgment and subadditivity: the role of working memory capacity and constraining retrieval. *Mem. Cognit.* 31, 968–982.

Fernandes, M. A., and Moscovitch, M. (2000). Divided attention and memory: evidence of substantial interference effects at retrieval and encoding. *J. Exp. Psychol. Gen.* 129, 155–176.

Fox, C. R., and Tversky, A. (1998). A belief-based account of decision under uncertainty. *Manage. Sci.* 44, 879–895.

Gettys, C. F., and Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organ. Behav. Hum. Perform.* 24, 93 –110.

Goldstein, D. G., and Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychol. Rev.* 109, 75–90.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychol. Rev.* 95, 528–551.

Juslin, P., and Persson, M. (2002). PROBabilities from EXemplars (PROBEX): a 'lazy' algorithm for probabilistic inference from generic knowledge. *Cogn. Sci.* 26, 563–607.

Kane, M. J., and Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: limits on long-term memory retrieval. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 336–358.

Luce, D. R. (1959). Individual Choice Behavior. Oxford, UK: John Wiley.

Libby, R. (1985). Availability and the generation of hypotheses in analytical review. *J. Account. Res.* 23, 646–665.

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

Mulford, M., and Dawes, R. M. (1999). Subadditivity in memory for personal events. *Psychol. Sci.* 10, 47–51.

Naveh-Benjamin, M., Craik, F. I., Guez, J., and Dori, H. (1998). Effects of divided attention on encoding and retrieval processes in human memory: further support for an asymmetry. *J. Exp. Psychol. Learn. Mem. Cogn.* 24, 1091–1104.

Patrick, J, Grainger, L., Gregov, A., Halliday, P., James, N., and O'Reilly, S. (1999). Training to break the barriers of habit in reasoning about unusual faults. *J. Exp. Psychol. Appl.* 5, 314–355.

Pleskac, T. J. (2007). A signal detection analysis of the recognition heuristic. *Psychon. Bull. Rev.* 14, 379–391.

Pleskac, T. J., Dougherty, M. R., Rivadeneira, A. W., and Wallsten, T. S. (2008). Random error in judgment: the contribution of encoding and retrieval processes. *J. Mem. Lang.* 60, 165–179.

Rohde, T. E., and Thompson, L. A. (2006). Predicting academic achievement with cognitive ability. *Intelligence* 35, 83 –92.

Rosen, V. M., and Engle, R. W. (1997). The role of working memory capacity in retrieval. *J. Exp. Psychol. Gen.* 126, 211–227.

Schooler, L. J., and Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychol. Rev.* 112, 610–628.

Sieck, W. R., and Yates, J. F. (2001). Overconfidence effects in category learning: a comparison of connectionist and exemplar memory models. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1003–1021.

Sprenger, A., and Dougherty, M. R. (2006). Differences between probability and frequency judgments: the role of individual differences in working memory capacity. *Organ. Behav. Hum. Decis. Process* 99, 202–211.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., and Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychol. Rev.* 115, 155–185.

Turner, M. L., and Engle, R. W. (1989). Is working memory capacity task dependent? *J Mem. Lang.* 28, 127–154.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131.

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

Unsworth, N., Heitz, R. P., Schrock, J. C., and Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37, 498–505.

Wallsten, T. S., Bender, R. H., and Li, Y. (1999). Dissociating judgment from response processes in statement verification: the effects of experience on each component. *J. Exp. Psychol. Learn. Mem. Cogn.* 25, 96–115.

Wallsten, T. S., and González-Vallejo, C. (1994). Statement verification: a stochastic model of judgment and response. *Psychol. Rev.* 101, 490–504.

Weber, E. U., Böckenholt, U., Hilton, D. J., and Wallace, B. (1993). Determinants of diagnostic hypothesis generation: effects of information, base rates, and experience. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 1151–1164.

Weber, E. U., Johnson, E. J., and Milch, K. F. (2007). Assymetric discounting in intertemporal choice: a query-theory account. *Psychol. Sci.* 18, 516–523.

Windschitl, P. D., and Wells, G. L. (1998). The alternative-outcomes effect. *J. Pers. Soc. Psychol.* 75, 1411–1423.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 April 2010; accepted: 02 June 2011; published online: 17 June 2011.

Citation: Sprenger AM, Dougherty MR, Atkins SM, Franco-Watkins AM, Thomas RP, Lange N and Abbs B (2011) Implications of cognitive load for hypothesis generation and probability judgment. Front. Psychology ***2****:129. doi: 10.3389/fpsyg.2011.00129*

This article was submitted to Frontiers in Cognitive Science, a specialty of Frontiers in Psychology.

APPENDIX

MODEL DESCRIPTION[1]

HyGene is implemented as a 6 step algorithm, which is initiated when the decision maker samples data from the environment and culminates with a set of potential hypotheses (and associated probabilities) that explain the observed data.

Step 1: data sampled from the environment (D_{obs}) initiates the activation of traces in episodic memory that represent past instances of the target hypothesis that share features with D_{obs}.

Step 2: the traces in episodic memory that are activated above a threshold value (A_{c}) enable the creation of an *unspecified probe* that is used as a retrieval cue to generate hypotheses from semantic memory (steps 3 and 4).

Step 3: the unspecified probe is matched against known hypotheses stored in semantic memory.

Step 4: hypotheses are generated from semantic memory and placed in the set of leading contenders (SOC) if they are sufficiently activated by the unspecified probe. The generation process involves sampling and replacement. Hypotheses are sampled from semantic memory according to their activation value. Recovered hypotheses are placed into the SOC if their activation values are greater than the least active member of the SOC. The SOC is a working memory construct and therefore is limited in capacity. Hypotheses in the SOC are referred to as "leading contender hypotheses."

Step 5: the posterior probability of each hypothesis in the SOC, $p(H_{\text{i}}|D_{\text{obs}})$, is given by comparing its memory strength to the memory strengths of all hypotheses in the SOC.

Step 6: hypotheses in the SOC are used to frame external information search or hypothesis testing.

The algorithm can be implemented in any number of representational systems. For our purposes, we use a system based on Hintzman's (1988) Minerva 2 model, and Dougherty et al.'s (1999) Minerva-DM model, and outlined in detail in Thomas et al. (2008) and Dougherty et al. (2010). In this model, events are represented as ordered sets of features, where values of +1, 0, or –1 are randomly assigned to each cell with equal probability (Hintzman, 1988). Traces in memory consist of three types of information that are relevant to modeling decision-making phenomena: data, hypotheses, and context. Encoding is modeled by a learning parameter, L. L determines the probability that each feature in the experienced event is encoded into the corresponding memory trace vector, where $0 < L < 1$.

Retrieval involves computing the similarity between a probe vector, P, and each trace T_i, in episodic memory, **M**. The similarity metric (S) used in HyGene is the dot-product between the probe vector and the trace, as defined by Eq. 2

$$S_i = \frac{\sum_{j=1}^{N} P_j T_{ij}}{N_i}, \tag{2}$$

Where, P_j is a feature in the jth position of the probe, T_{ij} is a feature in the jth position of the ith trace, and N_i = number of features where $P_j \neq 0$ or $T_{ij} \neq 0$. The activation, A**,** of trace i is given by cubing the value of S for each trace:

$$A_i = S_i^3. \tag{3}$$

HyGene assumes a conditional memory search process, wherein similarity is computed on only a subset of episodic memory, **M**, whose **K** traces contain data components sufficiently similar to the D_{obs}. Trace i is placed in the activated subset if its activation, A_i exceeds a threshold parameter, A_{c}:

$$A_i \geq A_c. \tag{4}$$

Traces included in the activated subset are probed a second time by the hypothesis, component of the probe vector, with the sum of the activations across the **K** traces in the activated subset giving rise to the conditional echo intensity:

$$I_{\text{C}} = \frac{\sum I_{A_i \geq A_c}}{K} \tag{5}$$

I_{C} is the mean conditional echo intensity and **K** is the number of traces for which $A_i \geq A_{\text{c}}$.

Hypothesis generation in HyGene begins by deriving a content vector from the subset of **K** traces activated by the initial retrieval cue, D_{obs}. The conditional echo content is a vector, $\mathbf{C}_c$, whose jth element is given by Eq. 6,

$$C_{\text{c}} = \sum_{i=1}^{K} A_i T_{ij}, \tag{6}$$

where $\mathbf{C}_{\mathbf{c}}$ is the conditional echo content for the jth element and **K** is the number of traces for which $A_i \geq A_{\text{c}}$. The echo content vector is normalized by the absolute value of the largest content value. This ensures that any positive content value greater than 1.0 and any negative content value less than −1.0 are perceived within the allowable feature range of +1 to −1, while preserving the sign of the original content values.

The conditional echo content process creates an *unspecified probe*. Hypothesis generation involves matching the unspecified probe against all known hypotheses in semantic memory in parallel and computing their activation values. Semantic memory employs a trace representation, but in contrast to episodic memory, semantic memory contains only a single representation of each hypothesis. Hypotheses in semantic memory whose semantic activation (A_{s}) is greater than zero define the semantic hypothesis space. Hypotheses are sampled probabilistically according to their activation value (c.f., Luce's choice axiom, Luce, 1959), and are recovered from semantic memory and added to the SOC if their A_{s} exceeds $\text{Act}_{\text{Min}H}$ (a rule that specifies the minimum activation necessary for a semantic trace to enter the SOC). Although the initial value of $\text{Act}_{\text{Min}H} = 0$, $\text{Act}_{\text{Min}H}$ is assumed to be dynamically updated based on the activation values of the hypotheses that have been generated from semantic memory. The generation process terminates when the *total* number or retrieval failures (the model fails to retrieve a novel hypothesis) exceeds TMAX (a retrieval parameter that determines how long the model searches semantic memory).

[1] For a full calculated example of how HyGene works, see the Appendix of Thomas et al. (2008).

Hypotheses in the SOC are ordered according to their activation values, with the member of the SOC with the highest A_s interpreted as the best explanation of D_{obs}. The working memory capacity parameter, ϕ, specifies the upper limit of how many hypotheses can be held in working memory.

Probability judgments for any given hypothesis are provided by comparing the memory strength for the to-be-judged hypothesis with the memory strength for alternative hypotheses in the SOC, as specified by Eq. 7,

$$P\left(H_i \mid D_{\text{obs}}\right) = \frac{I_{C_i}}{\sum_{i=1}^{w} I_{C_i}}. \tag{7}$$

Equation 7 normalizes the probabilities of the hypotheses within the SOC so they sum to 1.0. Excessive probability judgment and subadditivity arise when participants cannot maintain the full normative set of hypotheses in WM.