Psychological Review
2008, Vol. 115, No. 1, 155–185

0033-295X/08/$12.00 DOI: 10.1037/0033-295X.115.1.155

Diagnostic Hypothesis Generation and Human Judgment

Rick P. Thomas
University of Oklahoma

Michael R. Dougherty, Amber M. Sprenger, and J. Isaiah Harbison
University of Maryland

Diagnostic hypothesis-generation processes are ubiquitous in human reasoning. For example, clinicians generate disease hypotheses to explain symptoms and help guide treatment, auditors generate hypotheses for identifying sources of accounting errors, and laypeople generate hypotheses to explain patterns of information (i.e., data) in the environment. The authors introduce a general model of human judgment aimed at describing how people generate hypotheses from memory and how these hypotheses serve as the basis of probability judgment and hypothesis testing. In 3 simulation studies, the authors illustrate the properties of the model, as well as its applicability to explaining several common findings in judgment and decision making, including how errors and biases in hypothesis generation can cascade into errors and biases in judgment.

Keywords: hypothesis generation, probability judgment, subadditivity, working memory, diagnosis

Most tasks are not well structured but require the decision maker to impose structure on the problem so that a search for the solution can take place (Simon, 1973). Take the task of a clinical diagnostician as an example. The clinician's task is similar to that of a detective in that he or she can use clues (i.e., data) to generate possible explanations of the presenting symptoms and to search for additional clues (i.e., data) to test the explanations (Elstein & Schwarz, 2002; Weber, Böckenholt, Hilton, & Wallace, 1993). The clinician presumably generates likely diagnoses (i.e., disease hypotheses) and actively seeks information to evaluate the generated diagnoses. The clinician's search for information in the environment is likely guided by those diagnostic hypotheses he or she is presently entertaining. The data newly revealed through the search process are used to both evaluate the diagnoses currently under consideration and generate new diagnoses. The generation of new diagnoses may, in turn, lead to fresh information-search threads where new hypotheses might be brought to mind. At some point, either the search space is exhausted, the clinician continually fails to generate additional plausible hypotheses, or one particular hypothesis gains enough evidential support that the clinician can render a diagnosis with confidence (Elstein & Schwarz, 2002).

The example given above is only one of many real-world inductive inference tasks in which the decision maker is required to generate hypotheses on the basis of data. Indeed, hypothesis-generation processes occur in any task that involves taking data and formulating possible explanations of those data, including clinicians' generation of diagnoses on the basis of symptoms (Botti & Reeve, 2003; Elstein, Shulman, & Sprafka, 1978; Vermande, van den Bercken, & De Bruyn, 1996; Weber et al., 1993), auditors' diagnosis of going-concern problems on the basis of data obtained from accounting records (Libby, 1985), mechanics' diagnoses of auto failure on the basis of symptoms (Mehle, 1982), fault generation by expert power-plant operators (Patrick et al., 1999), scientists' interpretation of patterns of data (Fischhoff, 1977), and personality trait assessments generated on the basis of behavioral patterns (Zuckerman, Knee, Hodgins, & Miyake, 1995). Hypothesis-generation processes may even underpin reasoning and problem-solving processes in some clinical psychology patients, including the tendency of patients with schizophrenia to generate implausible (i.e., delusional) hypotheses in response to environmental stimuli and rumination in patients with depression. In all of these cases, the generation of hypotheses serves as the lynchpin for inductive inference, for evaluating the probability of various hypotheses, and for searching for information in the environment to test hypotheses. Obviously, probability estimation and hypothesis testing cannot be initiated until at least one hypothesis has been generated.

The purpose of this article is to describe a memory-based account of how decision makers generate and evaluate hypotheses. Our goal is to provide a general model of human judgment that describes how hypotheses are generated on the basis of data extracted from the environment, how the hypotheses generated from memory are used to make probability judgments, and how the generated hypotheses frame subsequent information search in hypothesis-testing situations. In the present article, we introduce our model, HyGene, and use it to describe how people

Rick P. Thomas, Department of Psychology, University of Oklahoma; Michael R. Dougherty, Amber M. Sprenger, and J. Isaiah Harbison, Department of Psychology, University of Maryland.

First authorship on this article is shared equally by Rick P. Thomas and Michael R. Dougherty.

This article is based on work supported by National Science Foundation (NSF) Grant SES-0624099 awarded to Rick P. Thomas and NSF Grants SES-0134678 and SES-0620062 awarded to Michael R. Dougherty. We gratefully acknowledge the many colleagues and collaborators whose feedback has helped shaped the ideas contained within this article, including Thomas Wallsten, David Huber, Eddy Davelaar, Jack Blanchard, Ana Franco-Watkins, Scott Gronlund, and the late Charles Gettys.

Correspondence concerning this article should be addressed to Rick P. Thomas, Department of Psychology, University of Oklahoma, Norman, OK 73019, or to Michael R. Dougherty, Department of Psychology, University of Maryland, College Park, MD 20742. E-mail: rthomas@psychology.ou.edu or mdougherty@psyc.umd.edu

generate hypotheses from long-term memory and how the generated hypotheses serve as input to probability judgment. We return to the issue of information search in the General Discussion.

To begin, we make the distinction between events external to the decision maker and events internal to the decision maker. External events represent what we call *the universe of possible states.* This is the exhaustive collection of events that are related (in some cases, causally) to data. We assume that the decision maker has internalized (through learning) some subset of the external events, and we use the term *hypothesis* to refer to one's mental representation of an external event.

The distinction between external events and one's mental representation of the external events (i.e., one's hypotheses) is represented by the two largest concentric circles in the Venn diagram in Figure 1: Nonrepresented events are denoted with a ?, indicating that these states of the universe are unknown to the decision maker. Although these unknown events exist in the ecology, they have not been experienced or registered by the semantic memory system. Represented hypotheses are denoted by H, which we take (for convenience) to be part of the semantic memory system. We assume that semantic memory is populated with representations of experienced hypotheses (H) and the data (D) with which they typically are associated. Hypotheses that are associated with overlapping data can be thought to form clusters, or "essentially similar" sets (cf. Venn, 1866, as cited in Kilinç, 2001). For example, a cluster might represent a set of diseases that share at least some symptoms. Clusters are represented in Figure 1 by hypotheses with common alphabetic subscripts. When prompted with a pattern of observable data (D_{obs}), we assume that people generate a set of leading contender (SOC) hypotheses, which are the decision maker's best-guess hypotheses concerning what gave rise to D_{obs}. The SOC is represented by the small nonshaded circle in Figure 1 labeled *Set of leading contender hypotheses in WM*. Because clusters of hypotheses tend to share data, the SOC will tend to be comprised of hypotheses from the same cluster, though cognitive limitations (working memory [WM] capacity) or task characteristics (time pressure) might prevent one from entertaining the exhaustive set of hypotheses. Our goal was to develop a model that capitalizes on semantic relatedness as a means of defining the reference class (e.g., the set of diseases that lead to sufficiently similar symptoms), while allowing the model to respect relative frequency information.

The above discussion highlights the three main questions addressed in this article. First, given a piece of observable data (D_{obs}), what are the potential psychological bases for how people define the set of possible explanations of D_{obs} (i.e., what is the most likely hypothesis, H, and what hypotheses populate −H)? Second, how do people generate possible explanations of D_{obs} from memory, and what cognitive factors affect how many and which hypotheses people generate (i.e., how do people generate the set of potential hypotheses that they actively entertain as possible explanations of D_{obs})? Third, what are the implications of the generation process for how people judge the probability of particular hypotheses?

These questions are addressed through the development and testing of a computational process model of hypothesis generation. Next, we provide a brief overview of the theoretical framework underpinning HyGene. This is followed by a review of the main empirical findings in the hypothesis-generation and probability judgment literatures that we wish to address with our theory. We then describe the computational details of the HyGene model and illustrate the properties of the model in three simulation studies. Our goal here is to advance a plausible model that describes how

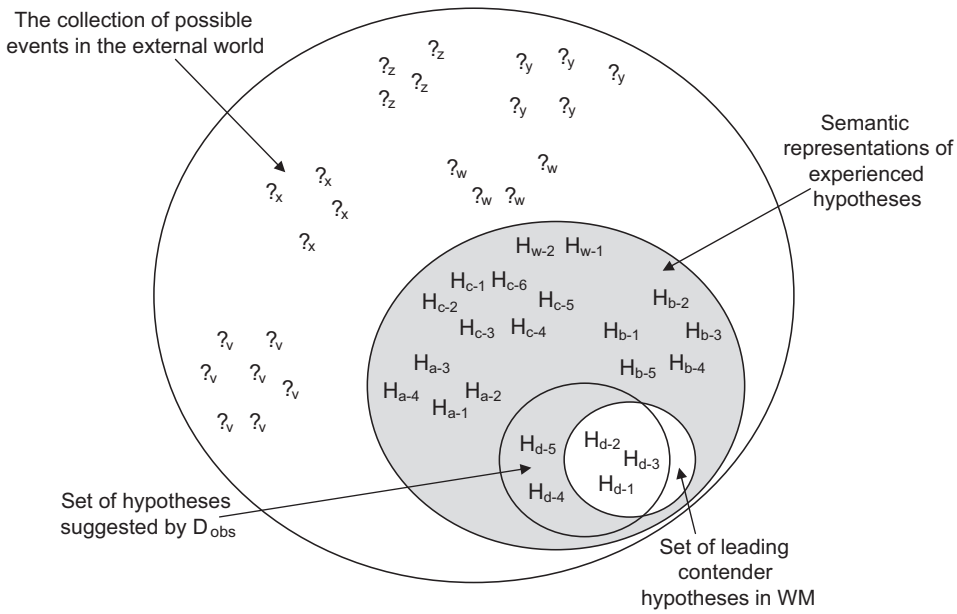

Figure 1. Venn diagram of semantics as defined within HyGene. Elements denoted with common alphabetic subscripts are members of the same cluster of hypotheses. Elements denoted by a ? are possible explanations of the D_{obs} that are not within the observer's semantic knowledge. Elements denoted with an H are within the observer's semantic knowledge. D_{obs} = pattern of observable data; WM = working memory.

humans generate sets of hypotheses, form probability judgments based on those sets, and select hypotheses from the generated set as explanations of observable data.

Theoretical Framework

HyGene is an integration of theoretical work in the areas of long-term memory, WM, and judgment and decision making and represents a major extension to Dougherty, Gettys, and Ogden's (1999) MINERVA-DM model. HyGene is based on three basic principles:

1. Data extracted from the environment serve as memory retrieval cues that prompt the retrieval of diagnostic hypotheses from long-term memory.

2. The number of diagnostic hypotheses that one can actively entertain at any point in time is constrained by both cognitive limitations and task characteristics.

3. Hypotheses maintained in the focus of attention (i.e., WM) serve as input into a comparison process to derive probability judgments and frame information search.

Principle 1 suggests that hypothesis-generation processes are a general case of cued recall in that the data or symptoms observed cue the retrieval of diagnostic hypotheses from either episodic long-term memory or knowledge. Note, however, that the retrieval goals in a hypothesis-generation task differ from the retrieval goals in the typical recall task. In many laboratory recall tasks, the goal is to retrieve a single exemplar or memory trace in response to a retrieval cue. Hence, the number of possible exemplars that could be considered a correct recall consists of a relatively small (and closed) set of possibilities, often consisting of a set size of one. In contrast, the retrieval goal in hypothesis-generation tasks is to generate a set of possible explanations for a given cue or set of cues. In most cases, a single cue (i.e., symptom) is related to multiple possible hypotheses, with the set of possible hypotheses often being indeterminately large (e.g., sore right abdomen is diagnostic of several conditions, including kidney infection, appendicitis, and bruising). That is, the set of possible alternative hypotheses that could explain (or cause) a particular cue (i.e., symptom) is ill defined.

In contrast to most recall tasks, the retrieval component in most judgment tasks is the first stage in a set of processes that require the decision maker to assess the probability of a set of possible hypotheses and/or to search for information that could be used to test (and potentially eliminate) the various hypotheses under consideration. Although the ultimate goal is to eventually arrive at the best explanation of the data, this process often requires that one evaluate and test multiple explanations to rule out alternative explanations.

Principle 2 merely states that the number of diagnostic hypotheses that a decision maker can actively maintain will be a function of WM limitations and constraints placed on the decision maker by task characteristics. For instance, HyGene assumes that the number of hypotheses that can be generated and held in the focus of attention is limited by WM capacity and is subject to dual-task constraints or divided attention (Dougherty & Hunter, 2003a). Because it takes time to populate WM with hypotheses, time pressure is assumed to lead to the generation of fewer hypotheses (Dougherty & Hunter, 2003b). For example, an emergency room physician likely will be forced to truncate the retrieval process if the patient's presenting symptoms require immediate action, such as resuscitation. Thus, time pressure is a type of task constraint that could affect the number of hypotheses considered (Dougherty & Hunter, 2003b). We assume that those hypotheses maintained in WM serve as input into the probability judgment and hypothesis-testing processes. Thus, probability judgments necessarily will be based on a subset of the possible hypotheses whenever the normative set of hypotheses exceeds the limits of one's WM capacity.

The third principle states that one's confidence in the generated hypotheses is determined by which hypotheses the decision maker is actively considering. Considerable theoretical and empirical work indicates that people make probability judgments by using a comparison process, where the strength of evidence for a focal[1] event is compared with the strength of evidence for a set of alternatives (Dougherty, 2001; Dougherty et al., 1999; Sprenger & Dougherty, 2006; Tversky & Koehler, 1994; Windschitl & Wells, 1998). The most well-known framework for describing this comparison process is support theory (Tversky & Koehler, 1994), which assumes that the probability of Hypothesis A, rather than Hypothesis B, is given by the evidential support for A, divided by the sum of the support for A and B: $p(A, B) = s(A)/[s(A) + s(B)]$. HyGene assumes that judgments of probability are derived in a similar fashion but specifies the cognitive processes underlying how hypotheses used in the comparison process are generated from memory and how the support values are assessed from memory.

The principles presented above form the basis of our theory of judgment. As we illustrate later through simulation methodology, the principles of generation from memory and cognitive limitations, coupled with a comparison process for judgment, go a long way toward accounting for a variety of judgment and decision-making phenomena. We argue that task characteristics, individual differences in WM capacity, and the ecological structure of the environment constrain hypothesis generation, which in turn leads to systematic effects on probability judgment.

Hypothesis Generation and Hypothesis Evaluation

Our development of HyGene was directed at explaining a set of findings in what we view as three interrelated literatures: (a) hypothesis generation, (b) hypothesis evaluation/probability judgment, and (c) information search/hypothesis testing. With few exceptions, research within these areas has proceeded relatively independently, with no real attempt to integrate findings across areas. As we argue below, we believe that the process of hypothesis evaluation/probability judgment is highly contingent on the hypothesis-generation process. Thus, by consequence, errors in the hypothesis-generation process will cascade into errors in probability judgment. In this article, we illustrate several properties of the HyGene model while addressing several empirical findings in the hypothesis-generation and probability judgment literatures. We provide an overview of the model's implications

[1] The term *focal hypothesis, focal event,* or just *focal* is reserved for the hypothesis the decision maker is making a judgment about. Thus, the focal could be a high-probability, low-probability, correct, or incorrect hypothesis.

for understanding hypothesis testing and information search in the General Discussion.

Hypothesis Generation

Although the evaluation of prespecified hypotheses has been the subject of research for many years, relatively little research has been concerned with the initial generation of the to-be-judged hypotheses (i.e., hypothesis generation). However, hypothesis generation arguably is more fundamental to human judgment than hypothesis evaluation and hypothesis testing. Indeed, it is the hypothesis-generation process that determines how many, and which, hypotheses are fed into the evaluation and testing processes.

To date there have been only a handful of studies concerning hypothesis generation. However, collectively, these studies have found four main results.

1. People generate relatively few hypotheses. Several studies have shown that people tend to generate only a subset of the total possible set of hypotheses (Gettys & Fisher, 1979; Gettys, Pliske, Manning, & Casey, 1987; Libby, 1985; Mehle, 1982; Weber et al., 1993). For instance, Mehle (1982) found that expert auto mechanics considered only four to six hypotheses or causes of auto failure when the proper set was considerably larger. Dougherty and Hunter (2003a) found that participants generated only three alternatives to the focal, even though they had learned through an exemplar-training task that the exhaustive set consisted of eight mutually exclusive hypotheses. Elstein et al. (1978) observed that expert physicians generated only about four alternative diagnoses prior to settling upon one. Finally, Dougherty, Gettys, and Thomas (1997) found that participants generated at most one or two alternative hypotheses when judging the probability of a particular causal scenario.

2. People tend to generate those hypotheses highest in a priori probability. A second finding in the hypothesis-generation literature is that participants tend to generate hypotheses that have the highest a priori probabilities. For example, Dougherty and Hunter (2003a) used a learning task and found that participants generated nearly twice as many high-base-rate hypotheses compared with low-base-rate hypotheses. A similar result was reported by Weber et al. (1993), who found that expert physicians regularly generated diagnoses that were highly likely given the descriptions of the hypothetical patients and nearly always generated the high-base-rate hypotheses prior to generating a low-probability (but high-cost) alternative. Similar results were reported by Dougherty et al. (1997) and Gettys et al. (1987). Thus, although people tend to generate only a subset of the plausible alternative hypotheses, those that are generated tend to have the highest a priori probability.

3. Number of hypotheses entertained is constrained by working memory limitations. We noted above that prior research revealed that participants generate fewer hypotheses than contained within the set of logical possibilities. However, one question is why the number of hypotheses considered by decision makers is so low. Perhaps not coincidentally, in all the studies cited above, the mean number of hypotheses generated by participants approximated the accepted range of WM capacity, 4 ± 1 (Cowan, 2001). More to the point, Dougherty and Hunter (2003a) found that a measure of WM capacity was positively correlated with the number of alternatives generated. In their study, high-span participants (top 25th percentile of their sample) generated roughly 4.1 hypotheses, compared with 2.4 hypotheses for low-span participants (the bottom 25th percentile of their sample).

4. People generate fewer hypotheses under time pressure. Dougherty and Hunter (2003b) examined, indirectly, the role of time pressure on hypothesis generation. Because the hypothesis-generation process involves the retrieval of hypotheses from long-term memory, they predicted that participants under time pressure, compared with participants without time pressure, would generate fewer alternative hypotheses and therefore give higher probability judgments. As predicted, judged probability was higher for participants under time pressure. This finding is consistent with recent research examining decision making in real-world tasks, where expert decision makers under time pressure often consider only a single hypothesis (cf. Flin, Slaven, & Stewart, 1996; Klein, 1993). Taken together, these studies suggest that the amount of time available to engage in hypothesis generation and/or time stress partially determine how many hypotheses will be generated and when the hypothesis-generation process will be terminated.

In summary, prior research has revealed that (a) participants generate fewer hypotheses than are possible, (b) participants tend to generate those hypotheses with the highest a priori probability, (c) participants are constrained in the number of hypotheses that can be explicitly considered due to WM capacity limitations, and (d) participants generate fewer hypotheses when placed under time pressure. As we indicate below, the number and a priori probability of the hypotheses considered will affect the perceived probability of the hypotheses under consideration.

Hypothesis Evaluation and Probability Judgment

One of the most studied behaviors in decision research has been how decision makers make probability assessments concerning prespecified hypotheses (i.e., hypothesis evaluation). Within this literature, several findings have emerged, many of which are accounted for by HyGene's predecessor, MINERVA-DM (Dougherty et al., 1999). Although our model can account for everything MINERVA-DM can explain, we focus on four additional findings that are outside the scope of the MINERVA-DM model.

1. Probability judgments tend to be subadditive. One robust finding in the probability judgment literature is that judgments tend to be subadditive: The judged probability of an implicit disjunction is less than the sum of the judged probability of its elements (Tversky & Koehler, 1994). As an example, Tversky and Koehler (1994) had participants rate either the implicit disjunction of p(death from natural cause) or had them rate three elements from the implicit disjunction p(death from heart disease), p(death from cancer), and p(death from other natural causes). The judged probability of p(death from natural cause) was .58, whereas the sum of the judgments to the three elements was .73. Clearly, participants' estimate of the implicit disjunction (death from natural cause) was less than the sum of the judgments for the explicit disjunction—a finding that has been replicated in a number of studies across a variety of tasks (Dougherty & Hunter, 2003a, 2003b; Dougherty & Sprenger, 2006; Koehler, 2000; Koehler, Brenner, Liberman, & Tversky, 1996; Mulford & Dawes, 1999; Sprenger & Dougherty, 2006; Tversky & Koehler, 1994).

2. Judged probability of an event is sensitive to the strength of its alternatives. Windschitl and Wells (1998; Windschitl & Young, 2001; Windschitl, Young, & Jensen, 2002) showed that the

judged probability of a focal hypothesis is sensitive to the strength of its alternatives, a finding they termed the *alternative outcomes effect*. Dougherty and Hunter (2003a, 2003b) extended this finding to show that the degree to which participants' probability judgments are subadditive is affected by the strength of the alternatives. For example, in one study, Dougherty and Hunter (2003a) found that both single-item probability judgments and the degree to which participants were subadditive were negatively correlated with the overall strength of the set of alternative hypotheses generated from long-term memory (Dougherty & Hunter, 2003a, 2003b). Importantly, the alternative outcomes effect demonstrates that the judged probability of any particular hypothesis is affected by the distribution of its alternatives.

3. Judged probability is related to hypothesis generation and working memory capacity. Dougherty and Hunter (2003a, 2003b) showed that probability judgments were negatively correlated with a measure of WM capacity and with the number of alternative hypotheses considered. In an extension of this study, Sprenger and Dougherty (2006) found a negative correlation ($r \approx -.25$) between WM span and judgments of probability but zero correlation ($r \approx .02$) between WM span and absolute frequency judgments. Sprenger and Dougherty argued that this finding reflected the fact that probability judgments, but not absolute frequency judgments, necessitate that the focal is compared with a set of alternatives. Accordingly, WM limitations are relevant only when the focal hypothesis must be compared with a set of alternatives because WM capacity constrains the number of alternatives to which the focal is compared. This finding is important because it demonstrates that the negative correlation between judged probability and WM span is not due merely to high-span participants' tendency to provide lower judgments.

4. Judged probability is greater under time pressure. As noted above, Dougherty and Hunter (2003b) found that judged probability was higher when participants made judgments under time pressure compared with no time pressure (see also Windschitl & Chambers, 2004). Dougherty and Hunter argued that time pressure truncated the amount of time participants had to generate alternative hypotheses. Consequently, judgments were higher because participants considered fewer alternatives under time pressure.

In summary, research has shown that participants' probability judgments tend to be subadditive, are sensitive to the distribution of the alternative hypotheses, and tend to increase as WM capacity decreases and as the amount of time available to generate alternatives decreases. As should be clear from our review, variables that have been shown to be related to the processes of hypothesis generation have been shown to have concomitant effects on probability judgment. In the next section, we present a computation model of judgment that accounts for the findings in the hypothesis-generation and the probability judgment literatures reviewed above.

HyGene

HyGene assumes three main memory constructs: (a) WM, (b) exemplar or episodic memory, and (c) semantic memory. WM is used for the maintenance of the SOC. The SOC is a subset of the total possible set of hypotheses that are maintained in the focus of attention. We assume that the SOC is limited by working memory capacity, which can be thought of as an individual-differences or task-constraint variable. Thus, one may maintain fewer hypotheses in WM because they have a relatively low WM capacity (an individual difference) or because their WM capacity is being consumed by a secondary task. In keeping with research within the WM literature, we assume that WM reflects one's ability to maintain goal-relevant information (i.e., hypotheses) in the focus of attention in the face of distraction (Engle, Tuholski, Laughlin, Conway, 1999).

Episodic memory is assumed to consist of a collection of traces that represents a database of the decision makers' past experiences. As such, this database preserves the experiential base rates of the various hypotheses within the decision maker's ecology, as well as a record of the data (or cues) that co-occurred with those hypotheses. This database represents the decision maker's internalized representation of the environment. The extent to which the probabilistic relationships between the hypotheses and data in the environment are maintained in the memory representation is referred to as *cognitive adjustment* (Gigerenzer, Hoffrage, & Kleinbölting, 1991). Events in memory are not perfect copies of the experienced events but rather are degraded copies, or traces, of the experienced events. The episodic memory representation is used as a means for extracting information from one's past experience that was systematically related to patterns of observed data (D_{obs}) in the environment. The episodic memory representation also is used for assessing the conditional probability of the various hypotheses (H) generated as possible explanations of D_{obs} (Dougherty et al., 1999).

In addition to episodic memory, HyGene also assumes a semantic memory.[2] Semantic memory is assumed to maintain both abstractions from the episodic system and generalized knowledge obtained outside of direct experience (e.g., book knowledge). Note that because semantic memory is based on abstractions, it lacks information about experiential base rates. That is, whereas a participant might have 1,000 traces of influenza and 20 traces of pneumonia in his or her episodic memory, both influenza and pneumonia are represented once and only once in the participant's semantic memory. Note also that the semantic system maintains representations of hypotheses learned outside of direct experience. For example, although one may have never seen a patient with malaria and hence have no record of malaria within the episodic system, he or she would be able to diagnose a patient with malaria because knowledge of malaria and its associated

[2] Considerable experimental and neuropsychological evidence in the memory literature supports the distinction between episodic and semantic memory (Mayes & Montaldi, 2001; Tulving, 2002; Tulving, Hayman, & Macdonald, 1991; Tulving & Markowitsch, 1998; Tulving, Schacter, McLachlan, & Moscovitch, 1988; Ward, 2003), and HyGene incorporates both types of memory systems to model hypothesis generation. It is nontrivial to model hypothesis generation using only an exemplar memory because the retrieval goals in most hypothesis-generation tasks differ from those of simple recall tasks. For example, in a cued-recall task, the goal of the participant is to retrieve a particular target item when cued with its associate: If one studies the pair *dog–tree* and is later presented with the cue *dog,* the goal is to retrieve *tree.* Hypothesis-generation tasks can be seen as an extension of the cued-recall paradigm with the goal of the participant being to generate multiple possible target hypotheses when prompted with the cue data. For example, physicians generate multiple diagnostic hypotheses when prompted with a symptom pattern (the data). This difference in retrieval goals, coupled with the possibility that identical or correlated data can occur with many different hypotheses, complicates the retrieval processes in such a way that retrieval based solely on the episodic memory system becomes intractable.

symptoms were learned in medical school and are represented in the semantic system.

Dougherty et al. (1999) conceptualized traces in memory as consisting of three types of information that are relevant to modeling decision-making phenomena: data, hypotheses, and context. For example, clinicians formulate diseases (hypotheses) from symptoms (data) and background patient information (context). Thus, we assume that both semantic memory and episodic memory represent information corresponding to hypothesis, data, and context information. For the sake of generality, the term *hypothesis* (H) is used to refer to anything about which the decision maker wants to make an inference (e.g., intentions, personality traits, causes, categories, explanations, solutions, stereotypes, diagnostic labels, and hypotheses about other symptoms or treatment options; Klayman & Ha, 1987). The term *data* (D) is used generically to refer to any piece of information that is used as the basis of inference (e.g., behavioral patterns, symptoms, or characteristics of the stimulus). The term *context* (C) is used to refer to any information that might be encoded as peripheral to the hypothesis and data. Data that is observed in the environment (i.e., a patient's symptomatology) is referred to as D_{obs}.

We assume that the three storage components enable one to populate the SOC with a set of plausible hypotheses through a prototype-extraction process and a semantic-activation process. Once the hypotheses have been generated from semantic memory into the SOC, they can then be fed into a comparison process that gives rise to overt probability judgments.

The prototype-extraction process involves the derivation of an *unspecified probe* from exemplar memory that is suggested by D_{obs}. The semantic-activation process involves a process of matching the unspecified probe against all known hypotheses in semantic memory to disambiguate it. Estimating posterior probabilities of hypotheses in the SOC (i.e., leading contenders) involves translating memory strengths into a probability/confidence judgment via a support-theory-like comparison process (Tversky & Koehler, 1994). Finally, hypothesis-testing strategies can be implemented to guide information search.

The basic structure of HyGene is illustrated in Figure 2. For ease of exposition, we use a medical diagnosis task to illustrate our model.

1. The process is initiated when data, D_{obs} (i.e., a symptom or a set of symptoms), are sampled from or observed in the environment. In the case of a clinician, these data might be the initial identification of a presenting symptom or a set of presenting symptoms. This initial sampling of data serves to initiate the activation of traces in episodic memory that represent past patients who have exhibited symptoms similar to D_{obs}.

2. The traces activated above a threshold value (A_c) ultimately result in the extraction, from episodic memory, of an unspecified probe that resembles those hypotheses that are most commonly (and strongly) associated with the data.

3. The unspecified probe is then matched against known hypotheses (e.g., prototypes of diseases) in semantic memory to determine the possible hypotheses for explaining the initial data D_{obs}, symptoms that are comorbid with D_{obs}, and potential treatments that have been associated with D_{obs} in the past.

4. Hypotheses in semantic memory are generated and can be placed in the SOC if they are sufficiently activated by the unspecified probe. The semantic traces are compared against the SOC for inclusion (or not) according to their activation (i.e., degree of match to the unspecified probe). The SOC is a WM construct and therefore is limited in capacity. Hypotheses in the SOC are referred to as *leading contender hypotheses* because they represent the decision maker's leading explanations for the presenting symptoms.

5. Once generation has ceased, the posterior probability of a particular hypothesis is evaluated by estimating the relative frequency of $H_i|D_{obs}$ in episodic memory. This evaluation process is accomplished via a conditional global-match process (cf. MINERVA-DM; Dougherty et al., 1999) where the judged probability of a particular hypothesis is given by its (memory) strength relative to the (memory) strengths of all hypotheses in the SOC. The output of this process is a conditional probability of $P(H_i|D_{obs})$ for each of the i hypotheses in the SOC.

6. The decision maker can engage in hypothesis-guided search by using hypotheses in the SOC to guide cue selection for hypothesis testing. We assume that diagnostic search can occur only when the decision maker is entertaining more than one hypothesis. Moreover, we postulate a consistency-checking process that eliminates leading contenders from the SOC that are inconsistent with D_{obs} (Fisher, Gettys, Manning, Mehle, & Baca, 1983). That is, we assume that a clinician rejects from the SOC any hypotheses that are inconsistent with the patient's symptoms.

The steps outlined above are assumed to be an iterative process, where the decision maker continually updates the SOC as new data are encountered in the environment and old hypotheses are rejected from the SOC. HyGene assumes that hypothesis-generation processes stop either when there is no time left or after the decision maker fails to generate new hypotheses on successive retrieval attempts (i.e., successive retrieval failures). Illustrative example calculations of HyGene that follow the six steps discussed above are provided in the Appendix.

Episodic Memory Processes in Hypothesis Evaluation

The episodic processes of HyGene build on the processes of MINERVA2 and MINERVA-DM. Although we specify the important components of HyGene's episodic processes here, more thorough treatments can be found in Hintzman (1988) and Dougherty et al. (1999).

Events in HyGene are represented as ordered sets of features. The set of features that specify an event is represented as minivectors. Minivectors can be filled or null. Memory processes ignore null vectors because the events do not exist in memory. Minivectors consist of N cells, where values of 1, 0, or -1 are

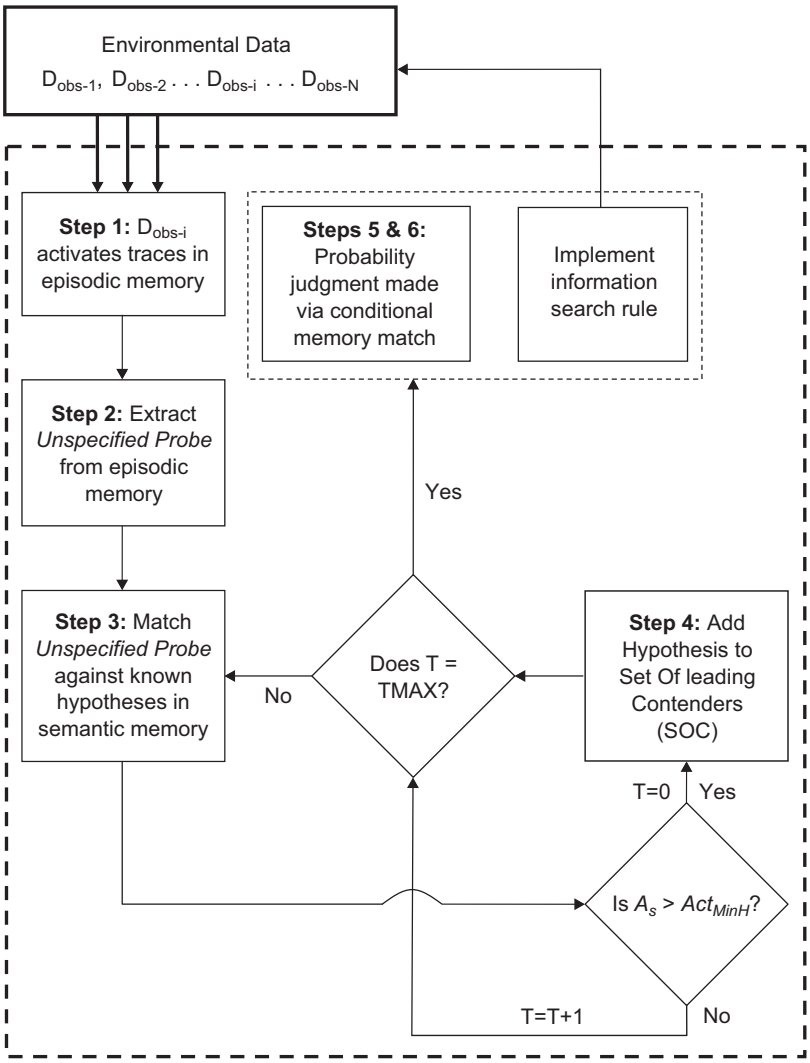

Figure 2. Overview of HyGene. A_s = activation of semantic $Trace_s$; Act_{MinH} = minimum activation threshold for hypothesis to enter the set of leading contenders; D_{obs} = pattern of observable data; T = number of consecutive retrieval failures; TMAX = parameter that defines the maximum number of consecutive retrieval failures allowed.

randomly assigned to each cell with equal probability (Hintzman, 1988). In our simulations, $N = 15$ for each minivector. We assume that events that occur together in the environment are represented by a single set of concatenated minivectors.

Encoding fidelity into episodic memory consists of creating copies (i.e., traces) of experienced events. The encoding fidelity or learning-rate parameter, L, specifies the extent to which a trace resembles the experience.[3] L is the probability that each feature of an event vector is encoded into the trace vector, where $0 \leq L \leq 1$. Degradation is modeled by converting a nonzero event feature (i.e., 1 or –1) into a 0 with probability 1 – L.

Retrieval is achieved by computing the similarity between a probe vector, **P**, and a set of traces in episodic memory. For simple retrieval tasks, it is assumed that a similarity value is computed for each trace, T_i, in memory, M. The similarity metric used in HyGene is the dot product between the probe vector and the trace. Similarity between a single trace, i, and the probe is given by Equation 1.

$$S_i = \frac{\sum_{j=1}^{N} P_j T_{ij}}{N_i} \quad (1)$$

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

[3] Note that as in MINERVA2 and MINERVA-DM, the model does not allow information to be encoded incorrectly (e.g., a 1 cannot be mistakenly encoded as a −1). However, this constraint is arbitrary and could be lifted if there was a need to model the effects of encoding errors on probability judgment or hypothesis generation.

$P_j \neq 0$ or $T_{ij} \neq 0$. The activation, A, of trace i is a positively accelerated function of the similarity between the probe and the trace:

$$A_i = S_i^3 \tag{2}$$

This cubing function serves as a weighting function, as those traces most similar to the probe will contribute more to the output of the model. Two types of information can be gleaned from matching episodic memory with the probe vector. One is called the echo intensity, I, and the other is the echo content vector, $\mathbf{C}$. I is the sum of the A_is across all M traces in memory:

$$I = \sum_{i=1}^{M} A_i \tag{3}$$

Equation 3 has been used to model both recognition memory and nonconditional frequency judgment. Dougherty et al. (1999) developed MINERVA-DM as an extension of MINERVA2 and used it to model conditional probability judgments, by postulating a two-part memory retrieval mechanism (i.e., conditional echo intensity). Imagine that a family clinician is asked to judge the probability of some hypothesis H (e.g., appendicitis) given some observed data, D_{obs} (e.g., sore right abdomen): P(appendicitis|sore right abdomen). We assume that the probability of H, appendicitis, is to be made conditional on D_{obs}, sore right abdomen, such that participants first partition episodic memory, M, into the subset of K traces that contain data components sufficiently similar to the D_{obs}. In our example, traces that contain data sufficiently similar to sore right abdomen would be placed in the activated subset. Trace i is placed in the activated subset if and only if the A_i between the D component of trace i and D_{obs} in the probe 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 H component (appendicitis) 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:

$$\mathrm{I_c} = \frac{\sum \mathrm{I_{A_i \geq A_c}}}{\mathrm{K}} \tag{5}$$

where I_C is the mean conditional echo intensity, and K is the number of traces for which $A_i \geq A_c$.

The second type of information that can be gleaned from episodic memory is information about its contents. Hintzman (1988) defined echo content as a vector of features consisting of the sum of the contents of all M traces, each weighted by its level of activation in response to the probe. Echo content is a vector, $\mathbf{C}$, whose jth element is specified by:

$$\mathbf{C_j} = \sum_{\mathrm{i=1}}^{\mathrm{M}} \mathrm{A_i T_{ij}} \tag{6}$$

Hintzman (1984, 1986, 1987) used echo content to model cued recall, as the content process enables the model to recover missing components based on the activation produced by the known portion of the probe (i.e., the cue). Hypothesis-generation tasks are similar to the cued-recall task in that both entail the retrieval of a trace in response to a cue. For example, imagine that if a clinician examined a patient with a sore right abdomen and probed memory with that symptom, the echo content returned would include sore right abdomen as well as events (i.e., other symptoms, disease states, etc.) that are associated with sore right abdomen in trace memory. Thus, hypothesis generation and associative recall are based on similar processes. However, the retrieval goal in associative recall is to retrieve a single item associated with the cue, whereas the retrieval goal in hypothesis-generation tasks is to retrieve a set of items (hypotheses) in response to a cue or pattern of data. Thus, HyGene extends the echo content process to allow for the modeling of hypothesis generation.

In keeping with the conditional processes in MINERVA-DM, we assume that a content vector can be derived from the subset of K traces activated by the initial retrieval cue, D_{obs}. We refer to this content vector as the *conditional echo content,* $\mathbf{C_c}$. The computation of conditional echo content differs from Hintzman's echo content calculation in two ways. First, we assume that the conditional content vector is based only on those traces in the activated subset. Second, the value of A_i used to compute content is based only on the cubed similarity between the specified portion of the retrieval cue (D_{obs}) and the corresponding component of the trace. This is the same value that is used to determine whether the trace is included in the activated subset.

The conditional echo content is a vector, $\mathbf{C_c}$, whose jth element is given by Equation 7.

$$C_C = \sum_{i=1}^{K} A_i T_{ij} \tag{7}$$

where $\mathbf{C_c}$ is the conditional echo content for the jth element and K is the number of traces for which $A_i \geq A_c$. Note that the vector $\mathbf{C_c}$ often will have content feature values outside of the allowable feature range of -1 to 1. Hintzman (1986) referred to this as the *ambiguous recall problem.* To solve this problem, 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 and any negative content value less than -1 are perceived within the allowable feature range of 1 to -1, while preserving the sign of the original content values.

The output of the conditional echo content and resolution processes is the creation of an unspecified probe. The unspecified probe contains information about events that have been associated with the symptom sore right abdomen in the past. Note that by conditionalizing on the subset of traces activated by D_{obs} (sore right abdomen), the model is in essence partitioning out the cluster of traces corresponding to hypotheses that are related to the known data. That is, events that have been systematically associated with D_{obs} (i.e., the observed symptom) in the past that are within the subset of activated traces are recreated in the conditional echo content vector. This property enables the model to function as a Bayesian inference engine and leads to an unspecified probe that is sensitive to the base rates of the hypotheses in the reference class defined by D_{obs}. The unspecified probe serves as the basis for the process of hypothesis generation.

Hypothesis Generation

The unspecified probe will represent the hypotheses in episodic memory that are related to D_{obs}. The determination of what the

unspecified probe might represent is achieved by matching it against semantic memory, which is assumed to maintain traces representing known hypotheses. Thus, the activation values between the unspecified probe and the known hypotheses stored in semantic memory determine which events in semantic memory are plausible interpretations of the unspecified hypothesis.[4]

Semantic memory employs a trace representation, but in contrast to episodic memory, semantic memory contains only a single representation of each hypothesis. For instance, we assume that a clinician could have multiple traces of patients with appendicitis in episodic memory but have only a single trace representing the prototype of appendicitis in semantic memory. Thus, semantic memory in HyGene takes the form of a list of prototypes. In the simulations, prototypes are modeled as the expected values of the relevant distribution of traces in memory. That is, the appendicitis prototype in semantic memory is the average of all appendicitis traces in episodic memory. Note that semantic memory lacks information about the base rates of the various hypotheses. However, because episodic memory retains one's experiential base rates and because these base rates are retained in the unspecified probe, HyGene's generation process is sensitive to the base rates of the various hypotheses. Note also, however, that the sensitivity to base rates is entirely predicated on the decision maker's experience as represented in episodic memory.

For simplicity, we assume that the unspecified probe activates all hypotheses in semantic memory in parallel. Hypotheses in semantic memory whose semantic activation (A_s) is greater than zero define the set of relevant hypotheses from which the decision maker is assumed to sample when generating hypotheses. Returning to Figure 1, this set is denoted by the circle labeled *Set of hypotheses suggested by* D_{obs}. The probability that a hypothesis is sampled from this set is determined by its activation relative to the activation of all other hypotheses in semantic memory with positive activation (cf. Luce's choice axiom; Luce, 1959). Hypotheses are generated from semantic memory and added to the SOC if and only if their A_s exceeds Act_{MinH} (a rule that specifies the minimum activation necessary for a semantic trace to enter the SOC), whose initial value is always set to zero. However, as discussed below, Act_{MinH} is assumed to be dynamically updated on the basis of the activation values of the hypotheses that have been generated from semantic memory.

Retrieval from semantic memory is assumed to terminate after TMAX consecutive retrieval failures, where a retrieval failure is defined by the failure to add a new hypothesis to the SOC (i.e., when $A_s \leq Act_{MinH}$) on a particular retrieval attempt. Because TMAX is a retrieval parameter that determines how long the model searches semantic memory, it can be used to model task characteristics, such as time pressure, and individual variables, such as effort or motivation (Webster & Kruglanski, 1994). Indeed, both time pressure and motivation have been shown to be important for determining how many hypotheses participants generate and the confidence they have for a chosen hypothesis (cf. Dougherty & Harbison, 2007; Dougherty & Hunter, 2003b; Kruglanski & Webster, 1996; Mayseless & Kruglanski, 1987; Webster, Richter, & Kruglanski, 1996).

Hypotheses in the SOC are assumed to be ordered by their overall resemblance to the unspecified probe, as given by their activation values A_s. Thus, the member of the SOC with the highest A_s is interpreted as being the hypothesis that is most similar to the unspecified probe. Note that this hypothesis may not be the hypothesis with the highest posterior probability (i.e., the best-guess hypothesis). In keeping with theory and research showing that WM is limited in capacity (Baddeley & Hitch, 1974; Engle, Kane, & Tuholski, 1999; Cowan, 1999), we assume that the number of hypotheses maintained in the SOC is dependent on one's WM capacity (Dougherty & Hunter, 2003a, 2003b; Sprenger & Dougherty, 2006). The WM-capacity parameter, ϕ, specifies the upper limit of how many hypotheses can be held in WM and can be used to model both individual differences in WM capacity (Engle, Kane, & Tuholski, 1999) and task constraints such as divided attention.

Technically, the hypothesis-generation process itself is not constrained by WM capacity. That is, over the course of solving a decision problem, a clinician might consider more hypotheses than can be maintained in WM. However, the number of hypotheses that can be held in WM at any point in time is capacity limited. For instance, if $\phi = 4$, then no more than four hypotheses can populate WM at any point in time for input into the evaluation and testing processes.

Dynamic updating of Act_{MinH}. We indicated previously that the value of Act_{MinH} is dynamically updated based on the activation of the hypotheses in the SOC. This assumption is based on work by Gettys and Fisher (1979), who found that participants' willingness to consider new hypotheses was dependent on the composition of the SOC. They demonstrated that as a decision maker generates hypotheses, his or her willingness to include a new hypothesis in the SOC decreases. Gettys and Fisher hypothesized that participants are willing to consider a new hypothesis if and only if it is a strong competitor of the most likely hypothesis currently under consideration.

The Act_{MinH} rule is parameter free in that its value is dependent only on the strength of the generated hypotheses, with the initial value being set at zero. Once the first hypothesis is placed in the SOC, its activation value sets the minimum activation level needed for subsequent hypotheses to be added. Thus, for a new hypothesis to be added to the SOC, it must have a higher activation than the first hypothesis that entered the SOC. Hypotheses continue to be added to the SOC so long as their $A_s > Act_{MinH}$, where $Act_{MinH} = \min(A_s \in \text{SOC})$. Once the number of hypotheses in the SOC reaches ϕ, new hypotheses that exceed Act_{MinH} replace the least active of the hypotheses in the SOC, and the value of Act_{MinH} is adjusted to reflect the minimum activation of the hypotheses in the

[4] The idea that semantic/lexical memory can be activated by episodic memory has parallels to models of text comprehension (see Caillies, Denhiere, & Kintsch, 2002; Kintsch, 2002; Kintsch & van Dijk, 1978; Rawson & Kintsch, 2002). Thus, HyGene assumes that the generation of known hypotheses takes the form of basic-level conceptual prototypes from semantic/lexical memory. HyGene assumes that these basic-level prototypes are derived from generalized episodic knowledge or book knowledge. For example, medical textbooks contain disease prototypes and production rules for diagnosis, and there is evidence that physicians rely on such abstracted knowledge when treating patients (Weber et al., 1993). The model of semantic memory implemented in HyGene is consistent with the idea that semantic categories are based on basic-level prototype representations (see Anderson, 1991; Hayes-Roth & Hayes-Roth, 1977; Reed, 1972). Although it is important to acknowledge that we are theoretically neutral regarding any particular representation of abstracted knowledge (e.g., production rules, templates, prototypes, inference networks, etc.), we do assume that this knowledge is learned.

updated SOC. Note that Act_{MinH} will increase incrementally as new hypotheses replace old hypotheses once ϕ has been reached.

The dynamic increase in Act_{MinH} has two by-products. First, the probability of generating a new hypothesis will decrease as the number of hypotheses that already have been generated increases. Second, the composition of the SOC is self-winnowing in that as the amount of time the model is given to search semantic memory increases, the composition of the SOC dynamically iterates toward containing the most likely hypotheses (i.e., hypotheses in semantic memory that have higher resemblances to the unspecified probe will displace ones with less resemblance) and the poorest hypotheses get eliminated from the SOC.

The dynamic updating of the Act_{MinH} criterion has important effects on the ultimate number of hypotheses generated. For instance, if hypotheses that weakly resemble the unspecified probe are generated initially, then ϕ is likely to be reached because Act_{MinH} will be relatively low. However, if the hypothesis that most resembles the unspecified probe is generated initially, it will be the only leading contender generated because no alternative hypotheses will have enough activation to pass Act_{MinH}. Moreover, highly likely hypotheses that most resemble the unspecified probe have the highest probability of entering the SOC through the interaction between the extraction of the unspecified probe and Act_{MinH}. Thus, the model predicts that when one dominating hypothesis exists, participants will generate only one hypothesis to explain D_{obs}.[5]

Consistency checking. Consistency checking is a process by which hypotheses are pared from the SOC by checking whether information or data associated with a generated hypothesis are consistent with observed data (Dougherty et al., 1997; Fisher et al., 1983; Patrick et al., 1999). Fisher et al. (1983) found that participants rejected, or eliminated from consideration, hypotheses in the SOC that were inconsistent with the available data. We assume that consistency checking is achieved by computing the similarity between the *i*th minivector in D_{obs} and the corresponding **D** minivector in the generated hypothesis. The decision rule for rejecting a hypothesis from the SOC is whether the S_i between the observed and hypothesized data is less than zero. Using zero as the criterion ensures that hypotheses are excluded from the SOC if and only if a datum in the hypothesis is highly dissimilar (i.e., has a negative similarity) to the data observed in the environment.

Hypothesis Evaluation/Probability Judgment

Probability judgments result from a comparison process that operates on the SOC. The judged probability of a focal hypothesis is given by its conditional echo intensity normalized by the sum of the conditional echo intensities for all leading contenders as specified in Equation 8.

$$P(H_i|D_{obs}) = \frac{I_{C_i}}{\sum_{i=1}^{w} I_{C_i}} \quad (8)$$

$P(H_i|D_{obs})$ is the probability of the *i*th hypothesis in the SOC, conditional on the subset of traces activated by D_{obs} (the data observed in the environment that were initially used to partition the subset of activated traces in episodic memory). The value of *w* in Equation 8 corresponds to the total number of hypotheses in the SOC, where $w \leq \phi$. Equation 8 recasts support theory's comparison process (Tversky & Koehler, 1994) in terms of echo intensities (i.e., memory strengths), where the support for the alternative hypothesis ($-$H) is given by the sum of the I_cs of only those hypotheses in the SOC.

HyGene's hypothesis-evaluation mechanism partitions the total probability among the explicitly considered hypotheses in WM (cf. Fox & Rottenstreich, 2003), with the partitions proportional to the values of I_c. Thus, HyGene's probability judgments will be additive for the set of explicitly considered alternatives. This property is referred to as *constrained additivity* and is an extension of support theory's property of binary additivity (for relevant data, see Dougherty & Hunter, 2003a). The assumption of constrained additivity requires that the judged probability of the focal decreases as the number and/or strength of alternatives in the SOC increase. Moreover, constrained additivity leads to the prediction that judgments will be excessive (and subadditive) when the number of hypotheses within the normatively exhaustive set of hypotheses exceeds *w,* the number of hypotheses explicitly considered by the decision maker. This is because the total probability is assumed to be partitioned over only those hypotheses explicitly considered (i.e., the SOC). However, if *w* equals the total number of possible hypotheses in the normative set, then probability judgments should be additive.[6]

The implications of hypothesis generation for probability judgment should be clear: Judgments are predicted to decrease as the number and/or strength of the hypotheses in the SOC increase. In the next sections, we present three simulations that detail the effect of several variables on hypothesis generation and probability judgments.

Simulation 1 was designed with two goals in mind. The first goal was to explore the relationship between several of the fundamental constructs within HyGene and its predictions. In particular, we examined the effect of encoding fidelity, experience, hypothesis base rate, and the similarity between the focal hypothesis and the alternatives. In so doing, we provided a demonstration of HyGene's sensitivity to variations in the parameters. The second goal was to examine the behavior of HyGene when instantiated as an ideal observer (IO) model. To this end, we developed an IO

[5] Act_{MinH} is just one example of a rule that satisfies the empirical constraints that a decision maker's willingness to admit new hypotheses to the SOC is partially dependent on which hypotheses are already being considered. There are many alternative rules, including Act_{AveH} (a rule that allows new hypotheses to enter the SOC if their activation exceeds the average activation of the leading contenders) and ones that would require parameterization (e.g., $Act_{MaxH}{}^{\gamma}$, a rule that allows new hypotheses to enter the SOC if their activation exceeds the activation of the strongest leading contender to the power defined by the parameter γ). We chose to implement Act_{MinH} in part because it did not require parameterization and in part because we had no a priori justification for choosing one rule over another.

[6] Note that this normalization process ensures that the probabilities in the SOC sum to 1.0, which means that the model outputs a probability of 1.0 for the focal hypothesis when it is the only thing generated. For consistency, we have implemented the comparison process (Equation 8) even when the model generates only a single hypothesis, but it is reasonable to assume that people respond with the raw I_c when only one hypothesis has been generated. I_c is a scalar value between 0 and 1.0. Given that I_c is based on the sum of the cubed similarities for all the traces that pass A_c, it reflects a compromise between the number of traces that pass A_c and their similarity to the probe. Thus, I_c can be interpreted as a valid probability.

model that optimized A_c, where we maximized the ability of the model to correctly discriminate between relevant and irrelevant episodic traces.

Simulation 2 was designed to examine HyGene's predictions regarding the effect of the strength of the alternatives and their similarity to the focal hypothesis on the degree to which judgments are subadditive (i.e., the degree to which judgments exceed 1.0) and also was designed specifically to model the results of Dougherty and Hunter (2003a). Simulation 3 was designed to examine the effect of time pressure and WM capacity on HyGene's subadditivity predictions.

Simulation 1

The primary purpose of Simulation 1 was to examine the implications of HyGene's first principle, which states that data from the environment act as a retrieval cue to prompt the generation of diagnostic hypotheses from long-term memory. In so doing, we explored how the model behaves across a variety of conditions, including the effect of encoding fidelity (modeled with the encoding parameter L), the amount of experience (modeled by manipulating the number of traces stored in HyGene's episodic memory), the similarity of the focal hypothesis relative to its alternatives, and the effect of relative frequency (i.e., base rates) of the hypotheses. The primary dependent variables were the total number of hypotheses generated by the model, the probability of generating and choosing the correct hypothesis, and the judged probability of the correct hypothesis. The correct hypothesis is the hypothesis whose data components are provided as the D_{obs}. The values of the parameters mentioned above are varied across a wide range, which enabled us to examine the flexibility of the model's predictions to changes in parameters. We implemented two versions of our model: a constrained version and an IO version. We elaborate on the differences between these two versions in the *Simulation Methodology* section.

A general prediction regarding the simulations is that the probability that HyGene will generate a particular hypothesis increases as a function of any variable that increases the memory strength of the hypothesis relative to its competitors. This relative increase in memory strength is predicted to result from both internal cognitive factors, such as better encoding of exemplars and increases in experience, and ecological factors, such as increases in relative frequency (i.e., ecological base rate) and the distinctiveness of the focal hypothesis. Several dependent measures were of interest. In all cases, the D_{obs} data used to probe episodic memory were causally related to the focal hypothesis. Thus, we were interested both in the number of hypotheses that HyGene generated into the SOC in response to D_{obs} and in the probability that the focal hypothesis (the correct hypothesis) would be selected. Finally, because we postulated that hypothesis-generation processes are important for probability judgments, we also were interested in examining HyGene's probability judgment predictions.

Simulation Methodology

The constrained model. For each simulated participant, eight different diseases were stored in semantic memory. Each disease consisted of 10 components (i.e., 10 filled minivectors). The first minivector denoted the disease component, whereas the other nine filled minivectors denoted the symptom components. Each minivector consisted of 15 features. Thus, each full disease vector was 10×15 features in length.

The alternative hypotheses were created to be more similar to each other than to the focal hypothesis. Figure 3 demonstrates how focal and alternative diseases were created. The prototype of the focal category and the prototype of the alternative disease category were randomly created to share *Sim* proportion of their features. The focal and alternative hypotheses shared .90 of their features on average with their respective category prototype.[7] Thus, the prototypes of the alternative hypotheses were more similar to the alternative category prototype than to each other. Also, the alternative prototypes were all more similar to each other than to the focal prototype. The similarity (parameter *Sim*) between the focal and the alternatives was systematically manipulated ($Sim = 0$, $Sim = .25$, $Sim = .45$, $Sim = .65$, $Sim = .85$, $Sim = 1$). Increasing the similarity between the focal and alternative hypotheses increases the strength of the alternatives to compete with the focal hypothesis to explain the data, making the focal and alternatives more confusable.

In many real-world tasks, the observed environmental data are not a perfect representation of the data associated with the causal hypothesis. That is, the data one perceives in the environment are often a perturbed or degraded version of the true underlying data. Thus, we assumed that the D_{obs} used to prompt retrieval was a degraded version of the true **D** vectors associated with the focal category prototype. This was modeled by using a D_{obs} probe that shared .85 of its features on average with the focal category prototype. The components of the data vector were presented simultaneously. This simultaneous presentation of the data can be contrasted with a serial presentation of the data in which the order that symptoms probe memory would affect the composition of the SOC. We describe a version of HyGene that deals with sequential sampling of cues in the General Discussion of the article.

To manipulate the relative frequency of the diseases (i.e., ecological base rates), each disease had traces in episodic memory distributed according to the distributions presented in Table 1. The first number of the distribution corresponds to the number of exemplars for the focal hypothesis. The focal hypothesis is the hypothesis that is causally related to D_{obs}. Distribution 1 represents a distribution in which the base rate of the focal hypothesis is .50. Distribution 2 represents an ecology in which the focal hypothesis has a relatively low base rate of .02 and the alternative hypotheses are quite prevalent (i.e., each alternative hypothesis is 7 times more prevalent than the focal hypothesis). Thus, the focal hypothesis has a relatively high base rate in Distribution 1 (.50) and a relatively low base rate in Distribution 2 (.02). However, the D_{obs} is always causally related to the focal hypothesis.

The effect of experience on hypothesis generation also was evaluated. Two levels of experience (E; low and high) were manipulated by multiplying the distributions in Table 1 by the constants 1 or 6. The traces were encoded with probability L,

[7] Note that both the D_{obs} (i.e., the probe) and the focal hypothesis (i.e., the correct hypothesis) shared .95 proportions of their features with the category prototype. Thus, the focal hypothesis shared .9025 proportion of its features with the D_{obs} because the *Sim* operations were independent ($.95 \times .95 = .9025$).

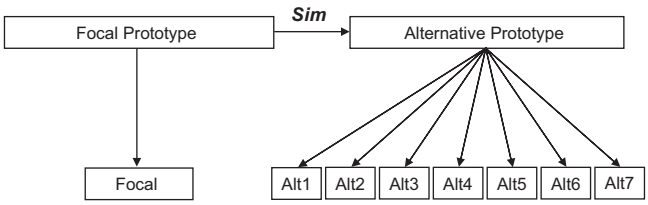

Figure 3. Semantic (similarity) structure of the hypotheses modeled in Simulation 1. Alt = alternative hypothesis; *Sim* = similarity parameter.

which was manipulated across six levels in the simulations (L = .1, L = .3, L = .5, L = .7, L = .9, and L = 1). All other HyGene parameters were held constant across the simulations (e.g., A_c = .216, $\phi = 5$, TMAX = 10, and the initial setting of $Act_{MinH} = 0$). Note that all parameters manipulated were fully crossed, so the simulations had a 6 (*Sim* = 0, *Sim* = .25, *Sim* = .45, *Sim* = .65, *Sim* = .85, *Sim* = 1) × 6 (L = .1, L = .3, L = .5, L = .7, L = .9, and L = 1) × 2 (E = 1 time, E = 6 times) × 2 (Distribution 1 = 70-10-10-10-10-10-10-10, Distribution 2 = 10-70-70-70-70-70-70-70) completely between-subjects experimental design. Thus, 144 HyGene Monte Carlo simulations of the constrained model were conducted for Simulation 1, with each simulation using 1,000 simulated participants.

Ideal observer model. We implemented an IO version of HyGene for model comparison purposes. There are several IO implementations possible within the framework. For the IO model, we chose to maintain all of the cognitive and ecological constraints present in the constrained model and decision ecology, but we optimized HyGene's conditional memory search parameter, A_c, within a signal-detection framework. The IO model represents a participant performing as if the A_c conditional memory search threshold were set optimally. Thus, the only difference between the IO model and the constrained model is the setting of A_c, which determines HyGene's ability to discriminate between set-relevant and set-irrelevant traces in episodic memory. The IO represents the upper boundary of performance of the HyGene model. The value of the optimized threshold (A_c) is computed so as to minimize product of (a) the ratio of the prior probability of traces that do not belong to the D_{obs} subset ($Trace_{D_{obs}=No}$) to the prior probability of traces that do belong to the D_{obs} subset ($Trace_{D_{obs}=Yes}$) and (b) a ratio of traces falsely identified as belonging to the D_{obs} subset (FA) and traces correctly identified as belonging to D_{obs} (Hits), as follows:

$$A_{C_{optimal}} \rightarrow f{:}\mathrm{Min}\left[\frac{P(Trace_{D_{obs}=No})}{P(Trace_{D_{obs}=Yes})} \times \frac{(FA)}{(Hits)}\right] \quad (9)$$

Note that we used Equation 9 to optimize the A_c for each simulation run or participant. As is seen in the simulation data, setting A_c optimally yielded an impressive ability of the model to recover the correct hypothesis (the one that gave rise to the data) even under conditions in which the correct hypothesis was highly confusable with its alternatives (high similarity value) and under conditions in which its base-rate probability was a mere .02.

Results and Discussion

The primary purpose of Simulation 1 was to examine how many hypotheses HyGene generates when prompted with cues (i.e., symptoms), the probability that the correct hypothesis is recovered and selected as the best guess, and the model's judged probability of the correct hypothesis. In the context of these simulations, the correct hypothesis is defined as the hypothesis that is causally related to the data used to probe episodic memory. The results of Simulation 1 are plotted in Figures 4, 5, and 6.

Number of hypotheses generated. It is instructive to examine the effects of encoding fidelity, similarity, and experience on the total number of hypotheses generated. As we argued in our review of the literature, the number of hypotheses people actively consider affects probability judgment.

Figure 4 plots the number of hypotheses generated by HyGene for each of the two distributions used in the simulations. Several results are noteworthy. One is that the number of hypotheses generated by HyGene is a relatively small subset of the eight total hypotheses in semantic memory and that on average, the size of this subset did not reach WM span, which was set at $\phi = 4$ for these simulations. Indeed, the finding that on average, the model generated only a subset of the total number of possible hypotheses is consistent with a variety of empirical results reviewed above (e.g., Dougherty et al., 1997; Dougherty & Hunter, 2003a; Gettys & Fisher, 1979; Libby, 1985; Mehle, 1982; Weber et al., 1993).

The failure of both the IO and constrained models to reach the maximum WM span on average is a result of the dynamic updating of the Act_{minH} criterion. This criterion has several interesting effects on hypothesis generation: First, because the value of Act_{minH} is based on the minimum activation of the hypotheses in the SOC, once one hypothesis has entered the SOC, Act_{minH} is incremented. Incrementing Act_{minH} prevents poorer contenders from entering the SOC, which in turn leads the model to generate a relatively small set of possible hypotheses. Furthermore, once WM span has been reached, Act_{minH} will only increase as new, better hypotheses enter the SOC and worse hypotheses are replaced. Thus, as Act_{minH} dynamically updates, the model becomes more selective in what it will allow into the SOC. This leads to the prediction that decision makers will be less willing to consider weak hypotheses as the number of generated alternatives increases. Moreover, within the context of studies examining the relationship between WM and hypothesis generation, the dynamic updating of Act_{minH} predicts a negatively accelerating relationship between the number of hypotheses generated and WM capacity. We return to this particular prediction for Simulation 3.

When *Sim* is low, the data are easily identifiable as belonging to the correct hypothesis. When *Sim* is high, however, the data from

Table 1
The Two Distributions of Traces Used in Simulation 1

Distribution	FH-A1-A2-A3-A4-A5-A6-A7
Focal prevalent condition	70-10-10-10-10-10-10-10
Alternative prevalent condition	10-70-70-70-70-70-70-70

Note: The numbers correspond to the number of traces for the focal hypothesis (FH) and each alternative hypothesis (A1, A2, A3, A4, A5, A6, and A7) stored in episodic memory for the low-experience condition. The higher experience condition was created by multiplying each trace frequency by 6. The focal prevalent condition corresponds to the case in which the FH is 7 times more likely than each alternative. The alternative prevalent condition corresponds to the distribution where each alternative is 7 times more frequent than the FH.

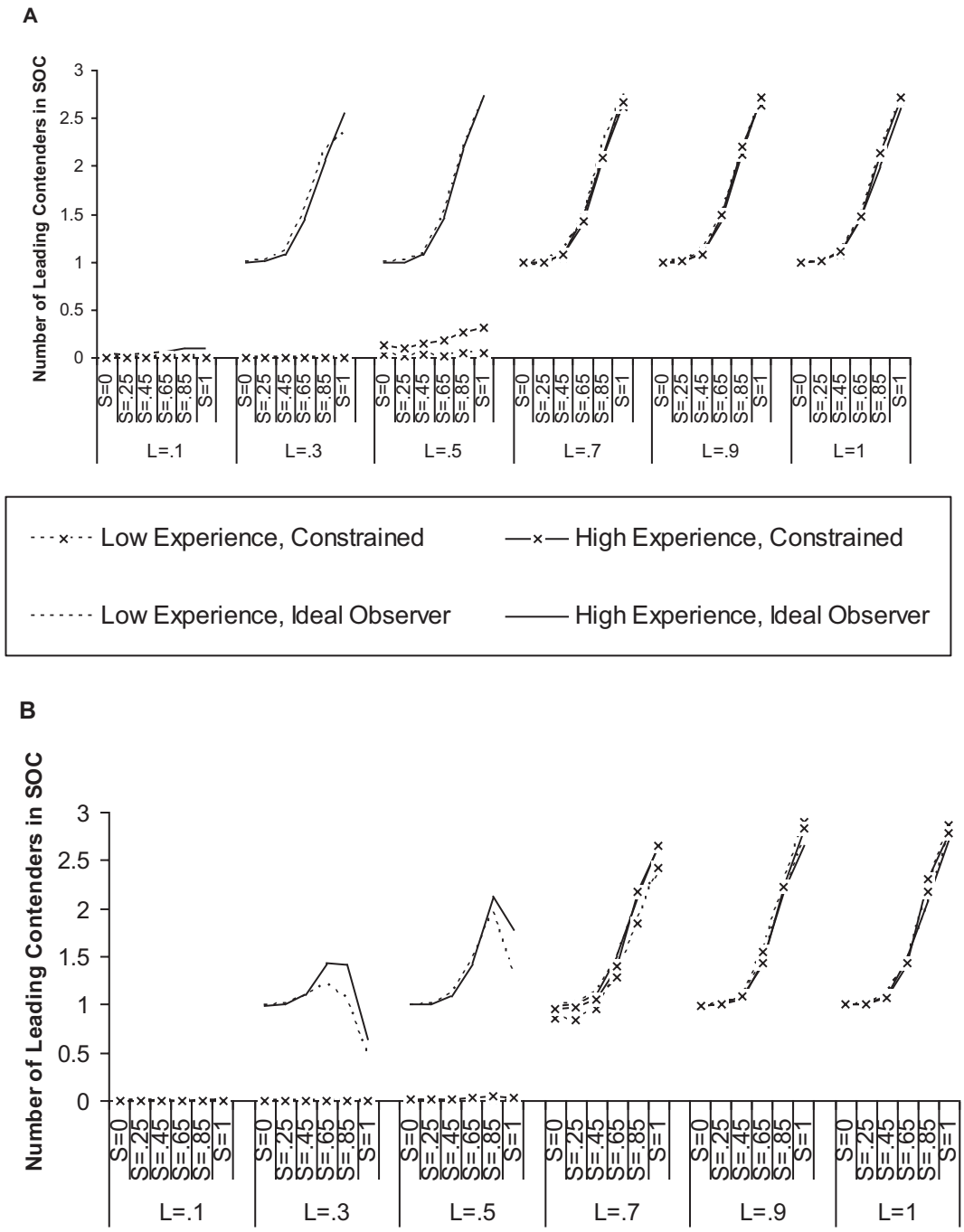

Figure 4. Average number of leading contenders generated as a function of distribution, similarity (S), and encoding fidelity (L). A: 70-10-10-10-10-10-10-10. B: 10-70-70-70-70-70-70-70. SOC = set of leading contenders.

the various alternative hypotheses are confusable with those of the correct hypothesis. Because the consistency-checking process prevents hypotheses that are highly dissimilar to the D_{obs} from being generated, as the similarity of the hypotheses decreases, fewer hypotheses pass the consistency-checking threshold. This leads to the prediction that the number of hypotheses generated depends on the similarity of the data among the various hypotheses in memory. For example, when *Sim* is relatively high (*Sim* $\geq$ .8), the model generates around three hypotheses. In contrast, when *Sim* is low (*Sim* = .5), the model generates only around one hypothesis. However, cases in which *Sim* is low are the conditions under which the model is also likely to generate the correct hypothesis even under conditions in which only one hypothesis is generated. Thus, although the model generates fewer hypotheses at low levels of

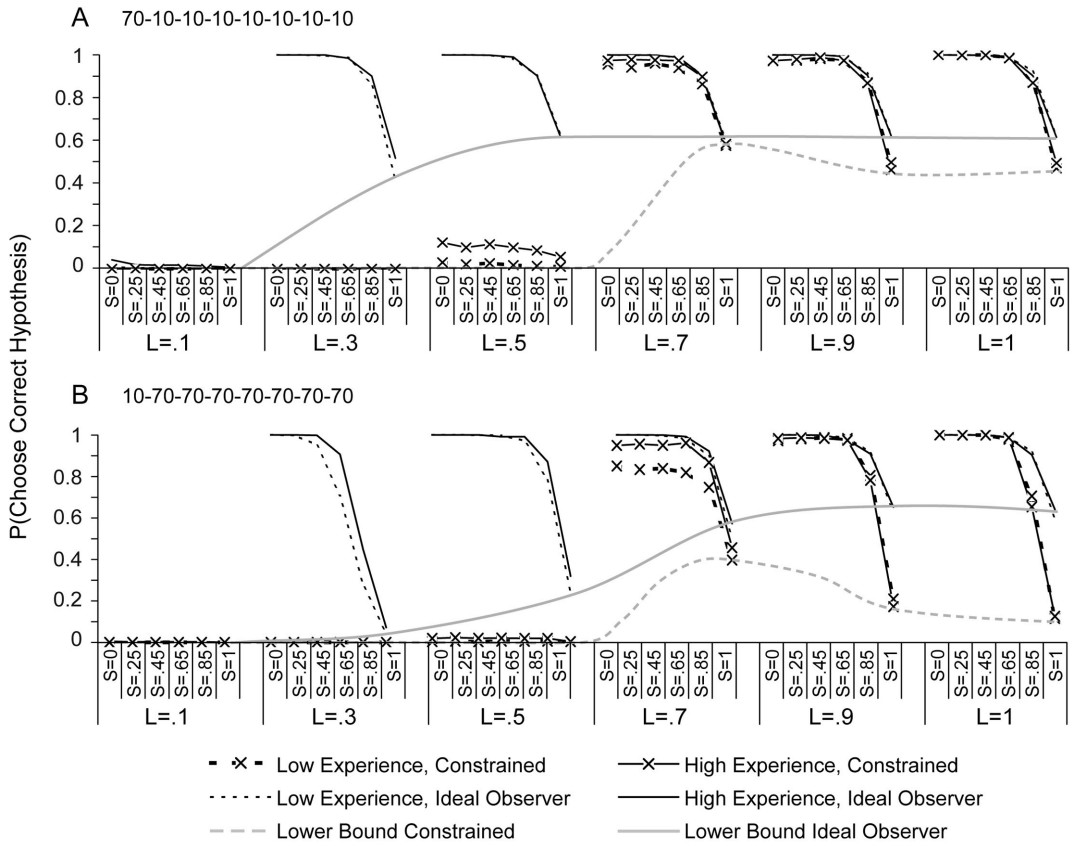

Figure 5. Probability (P) of choosing the correct hypothesis as a function of a priori probability (probe), similarity (S), and encoding fidelity (L). A: 70-10-10-10-10-10-10-10. B: 10-70-70-70-70-70-70-70.

Sim, the probability that the correct hypothesis is included in the SOC remains relatively constant because of the trade-off between number generated and the degree to which the data become more uniquely identifiable as belonging to the correct hypothesis.

Another interesting finding in Figure 4 is that there is little cost of modest decreases in L on the total number of hypotheses generated. For example, in Distribution 1, when the correct hypothesis is the high-base-rate hypothesis, the number of hypotheses generated by the model is nearly identical for L = .70, .90, and 1.0. Comparing this distribution with Distribution 2, where the correct hypothesis has a lower base rate compared with the alternatives, it is clear that the effect of decreased encoding fidelity depends on base rates and, to some extent, experience. As can be seen in Figure 4, the effect of experience on the number of hypotheses generated is modified by encoding fidelity (L): When encoding fidelity is high, there is little effect of experience, but when encoding fidelity is low, increased experience can offset poor encoding fidelity. In general, however, increased encoding fidelity leads to more hypotheses generated. The effect of increasing experience is most pronounced when encoding fidelity is poor (lower values of L) and when the correct hypothesis has a lower base rate than the alternatives. Note, however, that the number of hypotheses generated by HyGene is unaffected by experience and relative base rates when encoding fidelity is high. The IO model is able to generate more hypotheses at lower values of L than the constrained model due to the optimization of A_c. Moreover, the robustness of the IO model to poor encoding fidelity is stronger when the correct hypothesis is rare (Distribution 2) than when the correct hypothesis is prevalent (Distribution 1).

Probability of choosing the correct hypothesis. Although examining the model's predictions regarding the total number of hypotheses generated was important, it also was important to examine how often HyGene generated and chose the correct hypothesis. Figure 5 plots the mean proportion of times that the model chose the correct hypothesis.

There are several interesting findings shown in Figure 5. First, as encoding fidelity increases, the probability that HyGene chooses the correct hypothesis increases. One particularly interesting and nonobvious prediction demonstrated in Figure 5 is the three-way interaction between encoding fidelity, the similarity of the data (*Sim*), and distribution. In most cases, increased encoding fidelity leads to a higher probability of choosing the correct hypothesis. However, this is not the case when *Sim* = 1.0 (i.e., when the data associated with the correct hypothesis are highly similar to the data associated with the alternatives). In this distribution, the accuracy of the model actually benefits from information loss due to decreased encoding. This is demonstrated by the inverted U-shaped function between p(choose correct) and encoding fidelity when *Sim* = 1.0. Moreover, this effect causes the nonmonotonicity in the lower bound functions of the constrained model. Indeed, this effect

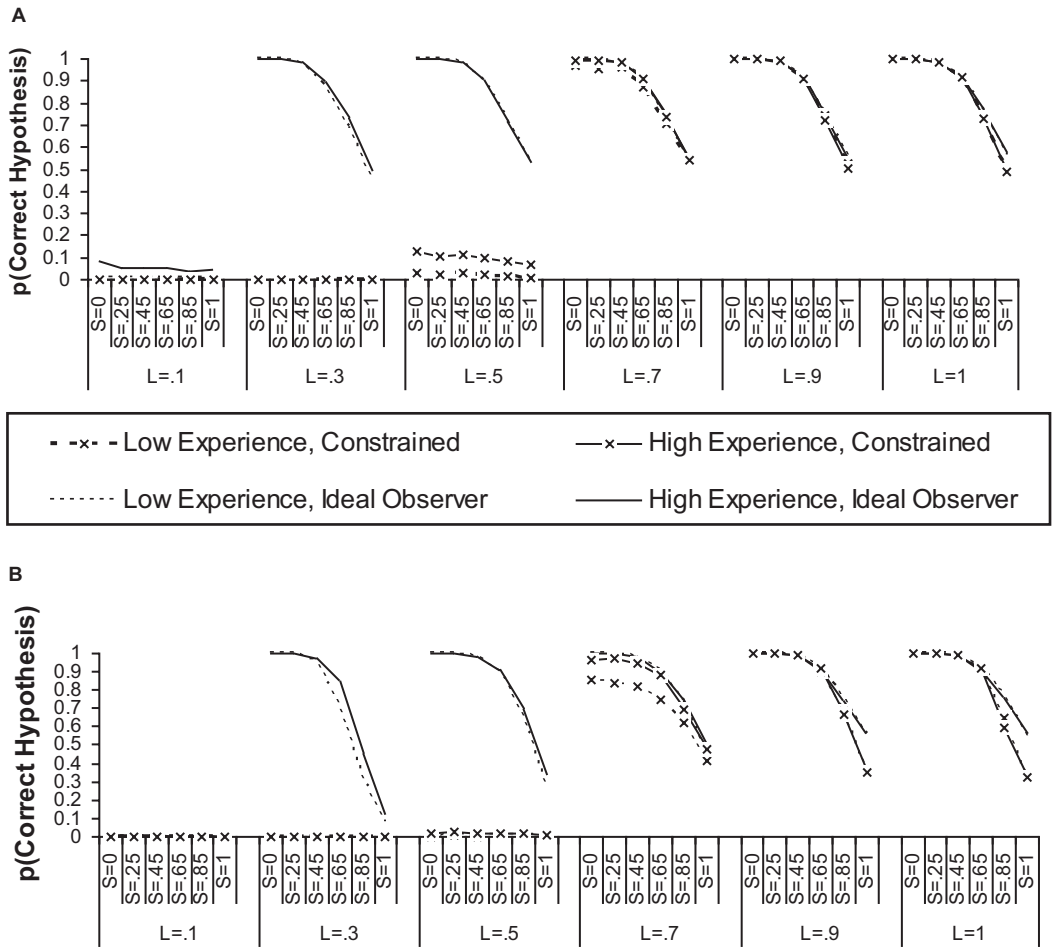

Figure 6. Probability judgment (p) of correct hypothesis as a function of a priori probability (probe), similarity (S), and encoding fidelity (L). A: 70-10-10-10-10-10-10-10. B: 10-70-70-70-70-70-70-70.

should be even more pronounced as the overall frequency of the alternatives increases relative to the correct hypothesis. Thus, HyGene predicts that imperfect storage, or loss of information due to forgetting, can be adaptive in some environments—most notably, when the relevant hypotheses in memory have a high degree of similarity to one another. Moreover, there is little cost, in terms of the probability of choosing the correct hypothesis, of modest decreases in encoding fidelity across both distributions (e.g., comparing L = .9 and L = 1.0).

Also of interest in Figure 5 is that HyGene is sensitive to hypothesis base rates: The probability of choosing the correct hypothesis is higher in Distribution 1, where the correct hypothesis is 7 times more frequent in episodic memory than each of the alternatives, than in Distribution 2, where each alternative is 7 times more frequent than the correct hypothesis. These findings indicate that the relative base rates of the hypotheses in episodic memory are preserved in the unspecified probe and affect the probability that the model will generate the correct hypothesis from semantic memory.

Note that the probability of choosing the correct hypothesis as shown in Figure 5 generally decreases with increases in *Sim*. As *Sim* increases, the model becomes less discriminating between the correct hypothesis and its competitors.

Judged probability of the correct hypothesis. In addition to examining HyGene's hypothesis-generation predictions, we also examined predictions regarding probability judgments. In the simulations, we examined predictions of probability judgments for the correct hypothesis regardless of whether it was generated by the model. Thus, if the model fails to generate a hypothesis or generates only the correct hypothesis, then the correct hypothesis is the only hypothesis in the SOC. Because of the constrained additivity assumption, the model assigns a probability judgment of 1.0 when the focal is the only hypothesis judged and at least one episodic trace is activated above threshold by D_{obs}. In contrast, if the model generates hypotheses other than the correct hypothesis, we assume that the correct hypothesis has been added to the SOC (i.e., to both the numerator and the denominator of Equation 8) because the model is prompted to judge it. In situations in which there is complete retrieval failure (i.e., no episodic traces are activated enough by the D_{obs} to pass threshold), the model renders a probability judgment of zero.

Figure 6 plots the mean predicted probability of the correct hypothesis for the two distributions as a function of experience, encoding fidelity, and *Sim*. Examining Figures 4 and 6 together, one can see that the judged probability of the correct hypothesis is negatively correlated with the total number of hypotheses gener-

ated—a finding supported in the literature (see Dougherty et al., 1997; Dougherty & Hunter, 2003a). The decrease in judged probability of the correct hypothesis with increases in *Sim,* as well as the decreases in judged probability with increases in L, is a consequence of the number of hypotheses generated and the strength of the hypotheses generated. In fact, plots where the model predicts asymptotically high probability judgments comprise those cases in which the correct hypothesis is the only hypothesis entered into Equation 9 and include those cases where the model fails to generate any hypotheses. Furthermore, as *Sim* increases, alternative hypotheses become stronger (as measured by conditional echo intensity) given the observed data. As the strength of the competitors increases, the judged probability of the correct hypothesis decreases because the strength of the correct hypothesis is compared with the strength of the competitors to derive probability judgment. Also note that there are systematic differences among the various distributions, as would be expected. For example, at high values of *Sim* and L, judgments of the correct hypothesis are higher in Distribution 1, where the correct hypothesis has a higher base rate than the alternatives, compared with Distribution 2. We more fully investigated HyGene's probability judgment predictions in the next two simulation studies.

The findings detailed above on the probability of choosing the correct hypothesis demonstrate that HyGene is behaving in a boundedly rational manner. Indeed, that the probability of the model generating the correct hypothesis is relatively robust to manipulations of *Sim* indicates that HyGene can deal with environments where the possible alternative hypotheses are good competitors (i.e., highly similar) to the correct hypothesis. Another demonstration of bounded rationality is the fact that the constrained model is performing at about the same level as the IO model. Simulations indicate that the probability of generating the correct hypothesis when only one hypothesis is generated remains relatively high despite poor encoding fidelity and high values of *Sim,* which suggests that HyGene can accommodate predictions based on Klein's (1993) recognition-primed decision-making theory and Gigerenzer and Goldstein's (1996) one-good-reason decision making. Although HyGene is irrational in the sense that it does not generate all possible alternatives, it is generating the most plausible alternatives given information-processing constraints, such as imperfect encoding fidelity, imperfect retrieval, limited WM capacity, and limited time to retrieve hypotheses from memory.

Importantly, Simulation 1 indicates how the HyGene cognitive architecture can be used to elucidate the relationship between memory and decision constructs and how understanding this relationship should be fruitful to judgment and decision making. Variables typically manipulated in decision research, such as base rates, as well as variables typically manipulated in memory research, such as encoding fidelity and experience, systematically affected the generation process. Thus, the behavior of the model indicates that HyGene's first principle (that data extracted from the environment serve as memory retrieval cues that prompt the retrieval of associated hypotheses from long-term memory) leads to hypothesis-generation behavior that is quite systematic. As we illustrate in the following simulation studies, the systematic effects of hypothesis generation have implications for hypothesis evaluation.

Simulation 2: The Alternative Outcomes Effect and Subadditivity

Two important findings in the probability judgment literature are the alternative outcomes effect and the subadditivity effect. The alternative outcomes effect is the finding that the magnitudes of participants' probability judgments are sensitive to the strength of the alternative hypotheses: Participants tend to give higher probability judgments when the alternatives are all relatively weak compared with when there are one or a few really strong hypotheses. A second finding in the probability judgment literature is that participants' probability judgment of a catchall hypothesis tends to be less (or subadditive) with respect to the sum of the probability judgments assigned to the individual hypotheses that make up the catchall hypothesis. Recently, Dougherty and Hunter (2003a, 2003b) revealed that the alternative outcomes effect and the finding of subadditivity are interrelated, in that the degree to which people are subadditive is affected by the strength of the alternative hypotheses. That is, participants show greater subadditivity when the hypotheses making up the catchall hypothesis are all relatively weak compared with when a few of the hypotheses are relatively strong.

Dougherty and Hunter (2003a) argued that the finding of subadditivity could be accounted for by assuming that participants generate hypotheses from long-term memory to use in a comparison process and that this comparison process is constrained by cognitive limitations. Furthermore, Dougherty and Hunter accounted for the finding that the degree of subadditivity is affected by the strength of the alternatives. That is, the covariation between the degree of subadditivity and distribution is due to including only a subset of the alternatives in the comparison process when estimating the probability of a particular hypothesis and to the fact that the subset of alternatives considered tends to be comprised of the strongest (i.e., most frequently occurring) hypotheses in the learned distribution. Coupled with the assumption of constrained additivity (i.e., total probability is partitioned over the set of explicitly considered hypotheses), this process predicts subadditivity whenever the decision maker fails to generate the complete set of alternative hypotheses. Moreover, this process also predicts that subadditivity should be less pronounced in distributions where the strongest alternative hypotheses have relatively higher frequencies (i.e., memory strength).

Simulation Methodology

The simulation methodology was designed to mirror the experimental methods used in Dougherty and Hunter (2003a). As such, for each of the 1,000 simulated participants, we created four different disease categories, with each category containing eight different disease hypotheses, for a total of 32 different diseases stored in semantic memory. Each disease category was defined by a unique symptom set, one for each distribution in Table 2. Table 2 presents the relative frequency with which the individual hypotheses within each category were represented in episodic memory. One can think of each distribution in Table 2 as characterizing different medical categories (e.g., blood disorders, cancers, psychological disorders, and immune disorders). Figure 7 provides a graphical depiction of how the eight diseases for each of the four distributions were derived. The prototype diseases for each cluster

Table 2
The Four Distributions of Traces Used in Simulation 2

Distribution	FH-A1-A2-A3-A4-A5-A6-A7
1	20-10-9-9-8-8-8-2
2	20-16-15-15-3-2-2-2
3	20-20-20-3-3-3-3-2
4	20-30-14-2-2-2-2-2

Note: The numbers correspond to the number of traces for the focal hypothesis (FH) and each alternative hypothesis (A1, A2, A3, A4, A5, A6, and A7) stored in episodic memory for each distribution. Note that in each distribution, the strength of the FH is the same (20 exemplars). However, the strength of the strongest alternative hypothesis (A1) increases, moving down from Distribution 1 to Distribution 4.

shared *Sim* proportion of their features, where *Sim* was manipulated at six levels (i.e., *Sim* = 0, *Sim* = .50, *Sim* = .80, *Sim* = .85, *Sim* = .90, and *Sim* = .95). Note that the eight diseases within each disease cluster were randomly selected to share .90 of their features with their respective category prototype. Each of the diseases consisted of nine data components and one hypothesis component (i.e., 10 filled minivectors). The WM-capacity parameter, ϕ, was manipulated at five levels (i.e., $\phi = 1$, $\phi = 2$, $\phi = 3$, $\phi = 4$, and $\phi = 5$). Parameters for this simulation were set at L = .95, TMAX = 10, and $A_c = .216$.

For each simulated participant, a probe, which shared .85 of its features with one of the disease category prototypes (e.g., blood disorders, cancers, psychological disorders, and immune disorders), was used as the retrieval prompt to generate hypotheses. We examined HyGene's predictions of the posterior probability of each of the eight hypotheses (i.e., diseases) within the disease cluster that was probed. For each judgment, HyGene was presented with a hypothesis vector associated with one particular disease, which prompted the model to assess a posterior probability using Equation 8. To derive the subadditivity score, we summed the model's predicted posterior probabilities for the eight hypotheses within the disease cluster (i.e., distribution) to which the probe was related.[8] Note that all parameters manipulated were fully crossed, so the simulations had a 6 (*Sim* = 0, *Sim* = .50, *Sim* = .80, *Sim* = .85, *Sim* = .90, and *Sim* = .95) × 5 ($\phi = 1$, $\phi = 2$, $\phi = 3$, $\phi = 4$, and $\phi = 5$) × 4 (distribution = 20-10-9-9-8-8-8-2, distribution = 20-16-15-14-3-2-2-2, distribution = 20-20-20-3-3-3-3-2, and distribution = 20-30-14-2-2-2-2-2) completely between-subjects experimental design. Thus, 120 HyGene Monte Carlo simulations were conducted for Simulation 2, with each simulation using 1,000 simulated participants.

Results and Discussion

Figure 8A presents the mean sum of the probability judgments from Simulation 2. For comparison, the results of Dougherty and Hunter (2003a) are presented in Figure 8B. As can be seen in Figure 8A, there was a clear effect of distribution on the sum of the probability judgments—an effect that mirrors the findings of Dougherty and Hunter (2003a). Note that the strength of the first hypothesis was kept constant across the distributions (i.e., Frequency 20 diseases), but the distribution of strength of the alternative hypotheses varied across the distributions. Thus, the differences between distributions are due entirely to the differences in the strength of the alternatives. Referring to Distributions 1–4 in Table 2, one can see that the stronger alternative hypotheses increased in strength from Distribution 1 to Distribution 4. The sum of the normalized conditional echo intensities of all the hypotheses (i.e., the amount of subadditivity; see Figure 8A) generally decreased as the distribution of strength of the alternative hypotheses became more asymmetric.

In model terms, this effect occurred because HyGene tends to generate the hypotheses with the highest relative frequencies (see Simulation 1). Moreover, because hypotheses with higher relative frequencies have relatively high conditional echo intensities compared with alternatives with lower relative frequencies, they tend to lower probability judgments of the correct hypothesis. This occurs because stronger alternative hypotheses receive a larger proportion of the probability space partitioning due to the constrained additivity property of HyGene's probability judgment mechanism. Thus, the simulation replicated the effect that the amount of subadditivity in probability judgments decreases as the number of traces of the strongest alternative hypotheses increases (i.e., the alternative outcomes effect).

Figure 8A shows that subadditivity decreases as the events being judged become more similar or confusable. As the similarity between the diseases increase, the alternatives become stronger

[8] In a single run of the model for Simulation 2, we prompted HyGene to generate hypotheses in response to D_{obs}. For illustration purposes, assume that the model generated H_1, H_5, and H_8 in response to D_{obs}. We wanted to obtain HyGene's posterior probability judgment for each hypothesis in the set of eight hypotheses and the sum of the eight posterior probability judgments to determine subadditivity. To illustrate, we first prompted the model to make a probability judgment about H_1, which we assumed, for the purposes of this example, was initially generated in response to D_{obs}. Here, the probability judgment would be the memory strength of H_1 divided by the sum of the memory strengths of all hypotheses in the SOC ($H_1 + H_5 + H_8$) via Equation 8. Next, we elicited the model's posterior probability judgment had we asked it about H_2 (as opposed to H_1), which was not initially generated by the model. Here, the probability judgment would be the memory strength of H_2 divided by the sum of the memory strengths of all hypotheses in the SOC ($H_2 + H_1 + H_5 + H_8$) including H_2 (the hypothesis prompted by the probability elicitation). To further illustrate, if we asked the model to judge the probability of H_3, which was not generated by the model, the probability of H_3 would be given by the memory strength of H_3 divided by the sum of the memory strengths of all hypotheses in the SOC ($H_3 + H_1 + H_5 + H_8$), H_3 being the hypothesis prompted (i.e., generated) by the probability elicitation and H_1, H_5, and H_8 being the hypotheses already residing in the SOC from the (self-)generation process. Thus, we were modeling eight what-if probability judgments on each run of Simulation 2 (and Simulation 3): what if we asked the model to judge the probability of H_1 in response to D_{obs}, what if we asked the model to judge the probability of H_2 in response to D_{obs}, what if we asked the model to judge the probability of H_3 in response to D_{obs}, what if we asked the model to judge the probability of H_4 in response to D_{obs}, what if we asked the model to judge the probability of H_5 in response to D_{obs}, what if we asked the model to judge the probability of H_6 in response to D_{obs}, what if we asked the model to judge the probability of H_7 in response to D_{obs}, and what if we asked the model to judge the probability of H_8 in response to D_{obs}? We then added the eight posterior probability judgments to derive the subadditivity score. Note that if we summed the probability judgments of the naturally generated hypotheses in our example, H_1, H_5, and H_8, they would sum to 1 due to HyGene's constrained additivity assumption. For an illustrative example, please see Step 6 in the Appendix.

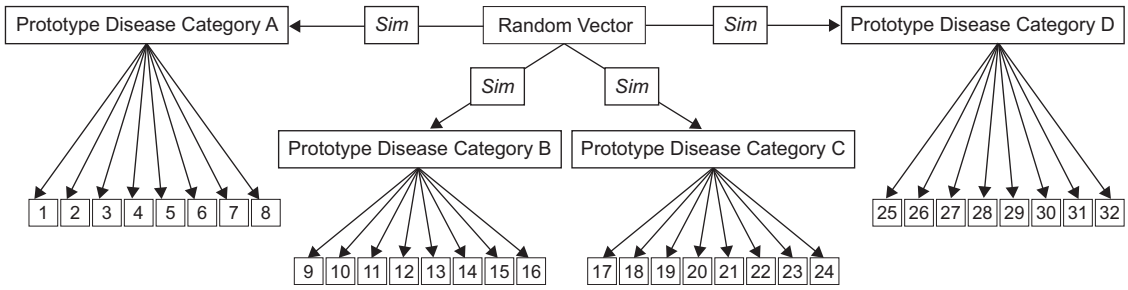

Figure 7. Semantic (similarity [*Sim*]) structure of the hypotheses modeled in Simulation 2.

competitors to the correct hypothesis, and subadditivity decreases as a consequence of HyGene's constrained additivity assumption (Equation 9). Figure 8A also indicates that subadditivity decreases as WM capacity increases. This replicates Dougherty and Hunter's (2003a) finding, depicted in Figure 8B, that high-span participants are less subadditive than low-span participants. The influence of WM capacity on probability judgment was explored in depth in Simulation 3.

Simulation 3: Effects of Working Memory Span and Time Pressure on Subadditivity

Dougherty and Hunter (2003b) argued that the finding of subadditivity could be accounted for by assuming that participants generate hypotheses from long-term memory to use in the comparison process and that this comparison process is constrained by cognitive limitations. In support of this finding, they found that the degree to which participants were subadditive was negatively correlated with a measure of WM capacity. In addition, subadditivity was found to be higher when participants made judgments under high time pressure, suggesting that time pressure truncated the hypothesis-generation process, leading to fewer hypotheses being included in the comparison process. In terms of HyGene, both WM capacity and time pressure affect the number of hypotheses included in the comparison process, as represented by w in Equation 8 and modeled with the WM-capacity parameter ϕ and time parameter TMAX. The purpose of this simulation was to examine the effects of WM capacity and time pressure on subadditivity. In so doing, we designed the simulations following the experimental paradigm used by Dougherty and Hunter (2003a, 2003b).

Simulation Methodology

For each of the 1,000 simulated participants, eight different diseases were stored in semantic memory. There was one randomly generated symptom set, and each disease shared an average of 90% of its features with the category feature set (i.e., prototype). Each disease consisted of 10 components (i.e., 10 filled minivectors). The first minivector denoted the disease component (i.e., hypothesis component), whereas the other nine filled minivectors denoted the symptom components (i.e., data components). Twenty exemplars of each disease were stored in episodic memory. The amount of time pressure was manipulated by varying TMAX across values of TMAX = 2, 4, 8, and 16. WM capacity, ϕ, was varied across values of ϕ = 1, 2, 3, 4, 5, 6, 7, and 8. All other HyGene parameters were held constant (e.g., L = .95, A_c = .216).

We asked HyGene to estimate the posterior probability of each of the eight hypotheses. To do so, episodic memory was probed with a degraded probe, where D_{obs} shared .85 of its features with the category feature set (i.e., prototype). For each judgment, HyGene was presented with a hypothesis minivector associated with one particular disease and was asked to assess the posterior probability using Equation 10. To derive the subadditivity score, we summed the model's predicted probability for the eight hypotheses.

Note that all parameters manipulated were fully crossed, so the simulations had a 4 (TMAX = 2, TMAX = 4, TMAX = 8, and TMAX = 16) × 8 (ϕ = 1, ϕ = 2, ϕ = 3, ϕ = 4, ϕ = 5, ϕ = 6, ϕ = 7, and ϕ = 8) completely between-subjects experimental design. Thus, 32 HyGene Monte Carlo simulations, each using 1,000 simulated participants, were conducted for Simulation 3.

Results and Discussion

Figure 9 presents the results of Simulation 3. Figure 9A plots the number of hypotheses generated as a function of WM span, ϕ, and TMAX, and Figure 9B plots the sum of the probability judgments as a function of WM span and TMAX. Several results should be clear. First, the number of hypotheses in the SOC increased monotonically with ϕ (see Figure 9A). However, note that the function is negatively accelerating and that the number of hypotheses generated is below ϕ, especially for higher values of ϕ. Thus, although ϕ sets the upper limit on the number of hypotheses generated, the model rarely reaches WM capacity. A second finding is the interaction between WM span and time pressure: There was no effect of ϕ among low values of TMAX, but as ϕ increased, the effect of TMAX on the number of hypotheses generated increased. This pattern of results suggests that the number of hypotheses generated by the decision maker is a function of both WM capacity and the amount of time spent generating hypotheses. Importantly, as we discuss next, these variables should have concomitant effects on judged probability.

Figure 9B plots the sum of the probability judgments for the eight mutually exclusive hypotheses as a function of ϕ and TMAX. As would be expected based on the number of hypotheses in the SOC, there was an effect of both variables on the sum of judgments. Judgments showed greater subadditivity under low values of TMAX and low values of ϕ. This finding was expected

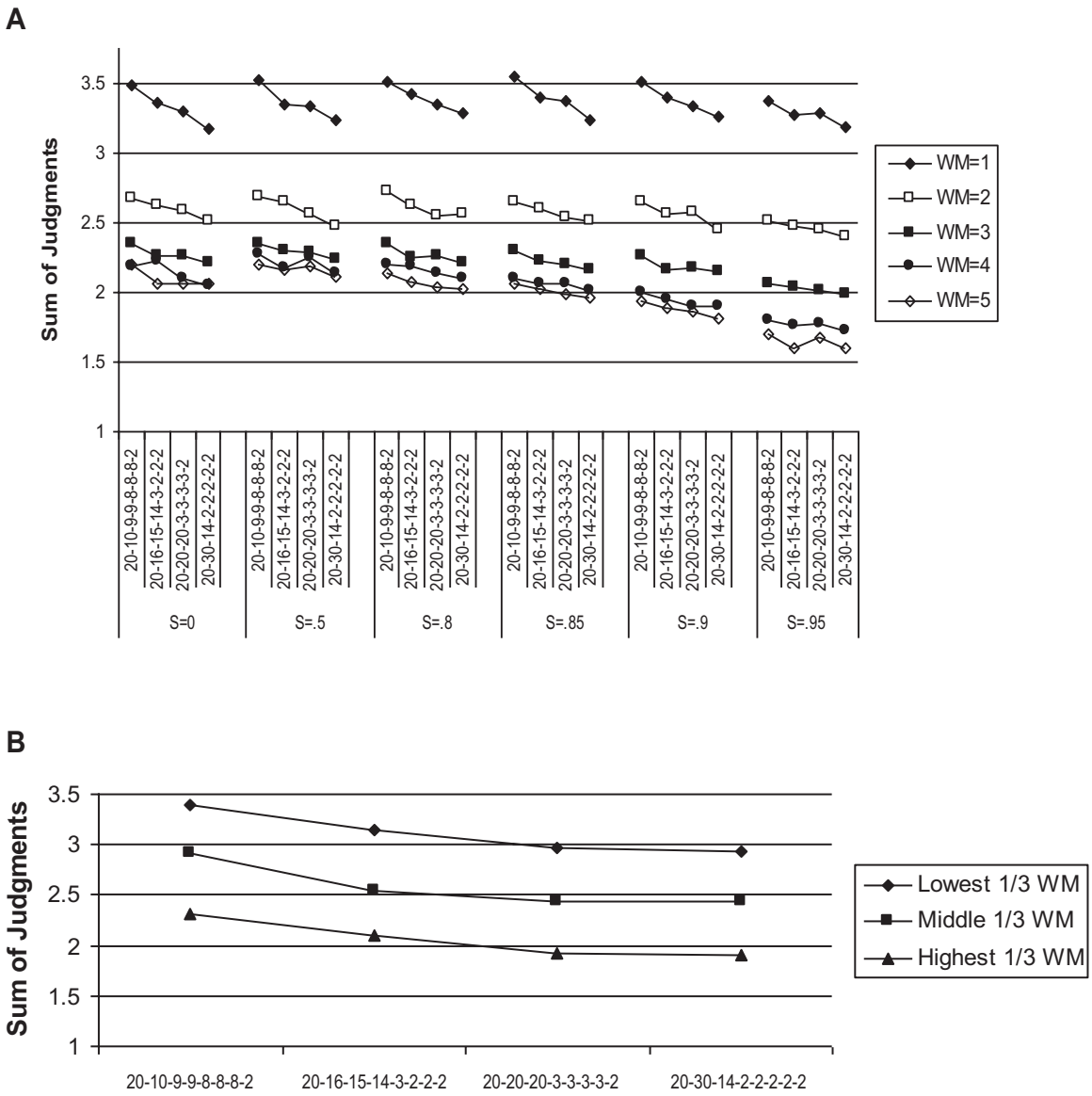

Figure 8. Results of Simulation 2. A: The sum of probability judgments assigned to the eight hypotheses within each distribution by similarity (S) and working memory (WM) span ($\phi = M$ [memory]). B: The sum of probability judgments by distribution and WM span. Panel B is reprinted from "Probability Judgment and Subadditivity: The Role of Working Memory Capacity and Constraining Retrieval," by M. R. P. Dougherty and J. E. Hunter, 2003, *Memory & Cognition, 31,* p. 975, with permission from the author.

because as ϕ decreases, the normalized conditional echo intensity or posterior probability judgment (i.e., $L[H_i|D_{obs}]$) of any particular hypothesis increases due to the constrained additivity property of HyGene's hypothesis-evaluation mechanism. That is, the probability space had to be partitioned over fewer leading contenders as both ϕ and TMAX decreased. Moreover, only under high values of ϕ was there an effect of TMAX on the model's probability judgments.

Note that the effect of both time pressure and WM span on the number of hypotheses generated is consistent with prior research. For example, Dougherty and Hunter (2003a) showed that the number of hypotheses generated by participants was lower for low-WM-span participants compared with high-span participants. Research in applied settings has revealed that decision makers placed under high time stress tend to generate relatively few hypotheses (Flin et al., 1996; Klein, 1993). Finally, there is evidence that increased time pressure leads to increased subadditivity (Dougherty & Hunter, 2003b).

General Discussion

The purpose of this article is to present a new theoretical framework that describes the cognitive processes underlying how people generate diagnostic hypotheses from memory and

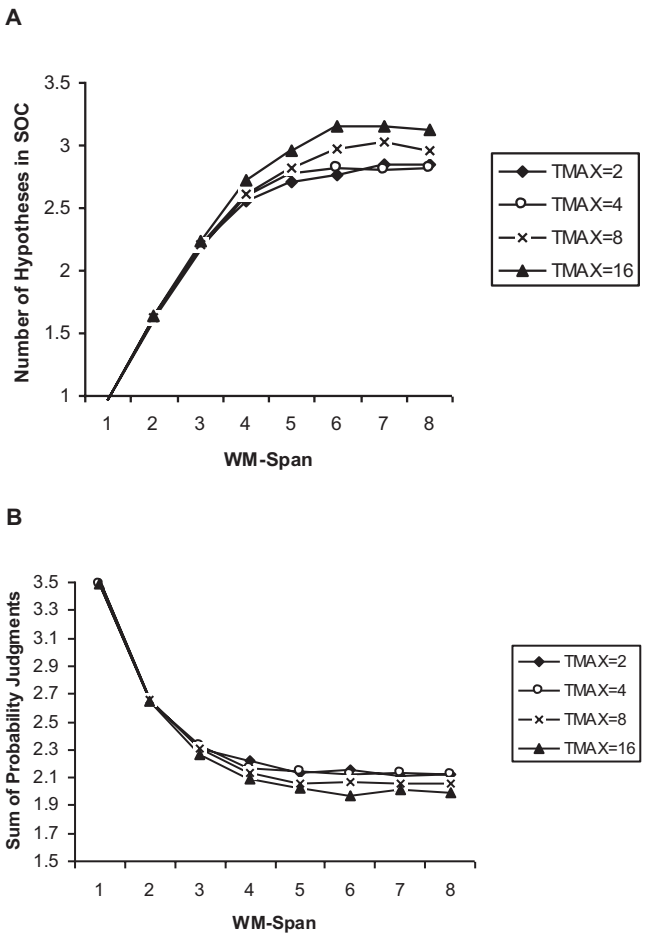

Figure 9. Results of Simulation 3. A: The number of leading contenders plotted as a function of working memory (WM) span ($\phi = M$ [memory]) and time pressure (TMAX). B: Average subadditivity plotted as a function of WM span ($\phi = M$ [memory]) and TMAX.

evaluate the probability of those hypotheses. As argued in the introduction, hypothesis generation is a fundamental component of human judgment. However, despite hypothesis generation's importance in understanding judgment, little empirical and even less theoretical work has been devoted to understanding the processes underlying hypothesis generation. As our simulations demonstrate, the generation of diagnostic hypotheses has important consequences for probability judgment and, as we argue below, also for understanding how people search for information to test hypotheses.

Our theory is based on three main principles: (a) Data extracted from the environment serve as memory retrieval cues that prompt the retrieval of diagnostic hypotheses from long-term memory, (b) the number of diagnostic hypotheses that one can actively entertain at any point in time is constrained by both cognitive limitations and task characteristics, and (c) hypotheses maintained in the focus of attention (i.e., WM) serve as input to the comparison process to derive probability judgments and are used to frame information search. We now turn to reviewing the main findings that bear on these three principles.

1. Data Extracted From the Environment Cues the Retrieval of Hypotheses From Long-Term Memory

Our first principle is that hypothesis generation is a general case of cued recall, in that data extracted from the environment (D_{obs}) serve as an initial retrieval cue to generate hypotheses from long-term memory. In terms of our simulations, one important factor in cuing long-term memory is the similarity of the D_{obs} to the various hypotheses in long-term memory. As Simulation 1 shows, the number of hypotheses generated from long-term memory is affected by how closely the various hypotheses in long-term memory resemble D_{obs}: The more uniquely identifiable the data were of the correct hypothesis, the fewer hypotheses the model generated. Importantly, the model was able to recover the correct hypothesis as part of the SOC roughly 70%–90% of the time regardless of the similarity among the hypotheses. This was true across a variety of values of encoding fidelity. One interesting finding is that the model actually benefited from a modest degradation in encoding fidelity when the hypotheses in long-term memory were highly similar to one another (*Sim* = 1.0). For example, when *Sim* = 1.0, the model performed better at retrieving the correct hypothesis when L = .70 than when L = 1.00, and this was particularly true when experience was high (6-times condition).

A third factor not presented in this article is that retrieval is highly dependent on the nature of the cue. For all of the simulations in this article, the probe (D_{obs}) was perturbed with 15% error. In simulation work not presented in this article, we found that the model's ability to recover the causally related hypothesis increases as D_{obs} more closely resembles the causal hypothesis but substantially decreases as D_{obs} is degraded. Thus, the behavior of the model suggests that the quality of D_{obs} partially determines a decision maker's ability to recover the casually related hypothesis from memory.

2. The Number of Diagnostic Hypotheses That One Can Actively Entertain at Any Point in Time Is Constrained by Both Cognitive Limitations and Task Characteristics

HyGene incorporates two constructs directed at modeling cognitive limitations: a WM-capacity parameter and a search time parameter. The WM-capacity parameter sets the upper limit on the total number of hypotheses that can be maintained in the SOC at any point in time. Interestingly, our simulations illustrate that the number of hypotheses generated often is less than the WM capacity (i.e., $w < \phi$). In particular, the model predicts that fewer hypotheses will be generated as D_{obs} becomes more uniquely associated with the correct hypothesis. As shown in Simulation 3, time constraints also are predicted to be important for hypothesis generation. Not surprisingly, the model predicts that more hypotheses will be generated as the amount of time available to generate increases. This prediction has been supported behaviorally by Dougherty and Hunter (2003b) using a probability judgment task that required participants to generate hypotheses from long-term memory. One interesting and nonobvious prediction also shown in Simulation 2 is that inducing time constraints reduces the advantage of having higher WM capacity. Thus, under conditions of high time pressure, HyGene predicts that there will be little difference between high- and low-span participants in the number of hypotheses generated.

One process that has not been explored is the idea that processing speed might figure prominently in hypothesis generation, particularly under conditions of high time pressure. A variety of studies have shown that processing speed is an important factor in long-term memory retrieval, especially cued- and free-recall tasks (e.g., Park, Smith, Lautenschlager, & Earles 1996). If one assumes a fixed amount of time for retrieval to take place, individuals with greater processing speed should be able to make more retrieval attempts within the fixed amount of time and therefore retrieve more alternatives. Thus, although capacity may not be an important individual difference under conditions of time pressure, individual differences in processing speed might be.

3. Hypotheses Maintained in the Focus of Attention (i.e., Working Memory) Serve as Input to the Comparison Process to Derive Probability Judgments and Are Used to Frame Information Search

All three simulations presented in this article illustrate the effect of hypothesis-generation processes on probability judgment. HyGene predicts that the judged probability of any particular hypothesis is a function of three factors: (a) the number of alternatives maintained in the SOC and included in the comparison process, (b) the strength (or objective frequency) of the hypotheses in the SOC, and (c) the similarity of the correct hypothesis to its alternatives. As our simulations show, both the judged probability of the correct hypothesis and the degree of subadditivity decrease as the number and strength of the alternatives maintained in the SOC increase. These two results have been found in a variety of behavioral studies, including Dougherty and Hunter (2003a) and Dougherty et al. (1997). The third prediction of HyGene is that the judged probability of the correct hypothesis also should be affected by the similarity of the correct hypothesis to its alternatives. This result has been produced with the MINERVA-DM model and subsequently supported by behavioral data (Bearden & Wallsten, 2004).

One aspect of the third principle not presented in this article is that the hypotheses in WM can be used to frame information search. At present, we are investigating several information-search heuristics that can be implemented within the HyGene framework. An important assumption in our investigation of these search heuristics is that hypothesis testing, or the search for information to test a hypothesis, is contingent on the SOC—an assumption we refer to as *hypothesis-guided search*. That is, we assume that people search only for information that is represented in the hypotheses that are maintained in the SOC. As a consequence, HyGene's information search follows a positive testing strategy when only one hypothesis has been generated but follows a diagnostic testing strategy when more than one hypothesis is being entertained by the decision maker. The memory-strength heuristic is an example of one search rule that can lead to positive test bias and the selection of pseudodiagnostic information. The memory-strength heuristic selects the cue that has the highest conditional echo intensity as calculated from the subset of traces activated by D_{obs}. It cannot discriminate whether that cue is actually diagnostic or not. As is probably obvious, when only one hypothesis is maintained in the SOC, the memory-strength heuristic selects positive tests and shows a preference for pseudodiagnostic over diagnostic information. In contrast, another search rule we are investigating, the similarity heuristic, chooses the cue that is most dissimilar between the hypotheses. The similarity heuristic can be implemented only when there is more than one leading contender. It never leads to the selection of nondiagnostic information.

Considerable research supports the idea that people engage in diagnostic search when at least one alternative is maintained alongside the focal hypothesis (Doherty, Chadwick, Garavan, Barr, & Mynatt, 1996; Kruglanski & Mayseless, 1988; Leblanc, 2003; Skov & Sherman, 1986; Slowiaczek, Klayman, & Sherman, 1992; Trope & Bassok, 1982; Trope & Mackie, 1987). Moreover, there is evidence that participants high in cognitive ability (a proxy for WM span) are more likely to engage in diagnostic search (Stanovich & West, 1998a, 1998b). Although no research has explicitly addressed the relationship between WM, hypothesis generation, and information search, the available evidence suggests the links are highly plausible.

Implications

The theoretical framework upon which HyGene is built has a number of implications, both theoretical and applied. We discuss five of these in turn.

The reference class problem and theories of probability. Venn (1866, as cited in Kilinç, 2001) argued that the probability of any particular event needs to be defined within a particular reference class, or set. The problem, as Venn noted, is that most, if not all, events can belong to multiple reference classes. This reference class dependence, often referred to as *the reference class problem,* means that one cannot define the true probability of a particular event, for the event will have a different probability when defined over different reference classes or sets. From Venn's perspective, sets were based on what he called *natural kinds.* We borrow Venn's idea that sets correspond to natural kinds but extend it by assuming that these sets are semantically defined in terms of shared data (or features). Natural kinds, in our rubric, merely correspond to sets of semantically related hypotheses.

To illustrate the reference class problem, consider Bayes's theorem, the cornerstone of classical decision theory. According to Bayes's theorem, the judged probability of a particular hypothesis is given by:

$$P(H|D) = \frac{P(H) \times P(D|H)}{P(H) \times P(D|H) + P(-H) \times P(D|-H)} \quad (10)$$

where H corresponds to the particular hypothesis under evaluation and –H corresponds to the collection of possible alternatives to H. A fundamental component of Bayes's theorem is knowledge of –H. Within a strictly normative framework, −H consists of all alternatives to H, and P(D|−H) is defined as the sum of probability of D for each element within −H.

Bayes's theorem provides a convenient description of how one ought to form probability judgments in well-defined environments where the entire set of elements contained within −H can be explicitly defined. Indeed, Bayes's theorem makes two crucial simplifying assumptions, namely, that the decision maker is provided the to-be-judged hypothesis and that −H consists of a well-defined set of alternatives. However, in real-world judgment tasks, decision makers are not always given the to-be-judged event (or the alternatives), and rarely is the set of possible hypotheses well defined. Moreover, a growing body of research indicates that

how people define −H (cognitively) leads to systematic violations of normative axioms. For example, Tversky and Koehler (1994), among others, showed that judgments are affected by the description of the events under consideration—a clear violation of the normative principle of descriptive invariance. More to the point, Dougherty and colleagues (Dougherty & Hunter, 2003a, 2003b; Dougherty & Sprenger, 2006) have shown that the perceived probability of any particular hypothesis depends on how many alternatives from –H are explicitly generated and included in the judgment process. Thus, regardless of the number of elements within –H, people are constrained in the number of elements that they actually consider in forming their judgments.

The simulations presented here offer insight into the cognitive processes underlying how people form reference classes and the basis of probability judgment. On the one hand, frequentists would argue that relative frequencies are the basis for probability judgment. On the other hand, subjectivists, as well as many cognitive psychologists, might argue that semantic relatedness carries relevant information that can inform judgments of probability and inductive inference. Regardless of which model one takes as normative, HyGene assumes that the human mind is adapted to respond to both semantic relatedness and relative frequencies. Within HyGene, semantic relatedness serves as a means for parsing semantic memory into clusters of essentially similar hypotheses, on the basis of which hypotheses share D_{obs}. The degree to which multiple hypotheses share D_{obs} determines, in large part, the size of the reference class. Inasmuch as D_{obs} is uniquely related to a single hypothesis (e.g., Simulation 1, $Sim = 0$), the reference class will be relatively small, with a lower bound of 1. However, the size of the reference class will increase as the number of hypotheses related to D_{obs} increases (e.g., $Sim \rightarrow 1.0$). Although semantic relatedness plays the major role in determining the reference class in semantic memory, the fact that probability judgments are based on exemplar memory ensures that HyGene is sensitive to relative frequencies.

HyGene's assumption that sets of hypotheses are represented as clusters in semantic space provides a computational approach to addressing the reference class problem—a problem that, some argue, besets most theories of probability (Hájek, 2007). Although HyGene does not resolve the philosophical debate surrounding the reference class problem, it offers a plausible model for how humans might form reference classes and the cognitive basis of probability judgment.

Artificial intelligence and decision support. Thus far, we have discussed HyGene as a model of human hypothesis generation and judgment. However, the model has obvious connections to work on artificial intelligence and decision support systems. One goal within the artificial intelligence literature is to develop systems based on models of human cognition that can outperform human observers. Our IO simulations illustrate such an approach.

The only difference between the IO model and the constrained model is the setting of A_c, which determines HyGene's ability to discriminate between set-relevant and set-irrelevant traces in episodic memory. Setting A_c optimally yielded impressive performance in terms of the model's ability to recover the correct hypothesis (the one that gave rise to the data) even under conditions in which the correct hypothesis was highly confusable with its alternatives (high similarity value) and under conditions in which its base-rate probability was only .02. Moreover, the IO model proved to be highly robust to degradation in how well items were encoded in memory: It performed equivalently across values of L ranging from 1.0 (perfect encoding fidelity) to .50 (where the episodic traces retained only 50% of the features present in the environmental events). Note also that the simulations utilized a degraded probe vector. So, the model was performing under conditions in which the data were error prone. Although Simulation 1 assumed a highly constrained environment with only eight hypotheses, these initial IO results provide a proof in concept that HyGene might provide a reasonable basis for developing cognitively inspired models of artificial intelligence.

The primary goal of decision support systems is to aid, not replace, human decision making (Shim et al., 2002). HyGene has the potential to play multiple roles within this field. First, it can provide a theoretical foundation to allow a better understanding of human decision processes, potentially supplying the necessary grounding to help developers of decision support systems to better anticipate and correct biases in human decision making (Arnott & Pervan, 2005). Second, as suggested by the IO version of the model, HyGene can itself serve as a decision support system. For example, it could be used to help in the generation of additional hypotheses. As discussed above, the initial structuring of the decision task is important if subsequent decisions are to be accurate. Clearly, to determine the potential successfulness of HyGene in this domain, much work is needed, such as testing it against alternative models (e.g., Stausberg & Person, 1999).

Clinical judgment. Elstein and Schwarz (2002) argued that clinicians use one of two types of reasoning processes to make diagnoses depending upon their expertise and on the difficulty of the diagnosis task. For simple diagnoses and diagnoses in which clinicians have high expertise, clinicians use pattern recognition processing. In these cases, physicians match a present case to a specific instance of a previous, similar case or to an abstract prototype. In these situations, Elstein and Schwarz argued, the clinicians do not engage in hypothesis testing. Rather, new cases are simply categorized by their resemblances to previous cases. For more difficult diagnoses, such as when more than one diagnostic category is activated (Weber et al., 1993), or tasks in which clinicians have less expertise, clinicians generate a limited number of hypotheses and use them to guide subsequent collection of data (hypothetico-deductive reasoning).

HyGene is consistent with this two-process conception of clinical judgment and specifies the conditions under which simple categorization processes will suffice and when more deliberative processes must be employed. For example, HyGene predicts that clinicians will rely primarily on pattern recognition when the to-be-judged patient's symptoms are highly similar to a single diagnostic hypothesis. As shown in Simulation 1, HyGene predicts that participants will generate a single hypothesis when the correct hypothesis is distinct from its logical competitors. However, under conditions in which the symptoms are related to multiple hypotheses, HyGene predicts that participants will generate multiple hypotheses. This is most clearly seen in simulations with high similarity values. Hypothetico-deductive reasoning processes are assumed to take place by the principle of hypothesis-guided information search, where hypotheses maintained in the SOC guide the search for new information that can be used to test or change one's belief about the set of hypotheses being considered.

In addition to providing a model of clinical judgment, HyGene also accounts for several findings in the clinical judgment literature. For example, considerable research indicates that physicians generate a small subset of the total number of alternative hypotheses, usually consisting of about four alternatives, even when the total set of potential hypotheses is much larger (Barrows, Norman, Neufeld, & Feightner, 1982; Elstein et al., 1978; Joseph & Patel, 1990; Weber et al., 1993). According to HyGene, this limitation in the number of hypotheses generated can arise from several different sources: WM limitations, motivation, time constraints, or the structure of the ecology. WM limitations place an upper boundary on the number of hypotheses that one can explicitly maintain. However, WM limitations will have little effect under conditions of high time pressure or low motivation (in which case, the decision maker may truncate search after generating one or two hypotheses) or when the symptoms are highly similar to one and only one hypothesis (i.e., when there are no competitors that share D_{obs}).

Weber et al. (1993) found that physicians generated high-base-rate diagnoses more often than other diagnoses. This finding is consistent with HyGene's prediction that hypothesis generation will be affected by base rates, as demonstrated in Simulation 1. There, the probability of generating the correct hypothesis was higher in Distribution 1, where the focal was 7 times more frequent in episodic memory than each alternative, than in Distribution 2, where each alternative was 7 times more frequent than the focal in episodic memory.

Finally, HyGene predicts that the probability of generating the correct hypothesis within a clinical session will be highly dependent on the initial set of generated hypotheses (Barrows et al., 1982). Within HyGene, the initial set of hypotheses that are generated has two consequences. First, through the operation of Act_{MinH}, the initial set of hypotheses sets the minimum criterion needed for new hypotheses to be added to the SOC. Thus, if the correct hypothesis is not included in the SOC early on, it is unlikely to be generated. This finding is consistent with Barrows et al.'s (1982) finding that the failure to generate the correct hypothesis within the first 30 seconds of a clinical interview often results in an incorrect diagnosis. Second, because the hypotheses in the SOC guide information search, the search for data for testing the hypotheses under consideration will likely exclude data directly related to the correct (ungenerated) hypothesis.

Still, there is much work needed before HyGene captures the dynamics of real-world clinical judgment. For example, thus far, we have used rather sparse hypothesis spaces for our simulated ecologies. For example, our largest simulation, Simulation 2, included only 32 hypotheses (four clusters with eight hypotheses each). In contrast, Gordon (1970) estimated the number of diseases to be approximately 6,000 and the number of symptoms to be approximately 20,000. Second, clinicians generally observe symptoms sequentially over time, rather than simultaneously. Thus, HyGene will need to be extended to model the sequential acquisition of data before it can be used as a model of physician hypothesis generation. The development of such a model is not trivial and requires, among other considerations, that we understand both how people maintain the generated hypotheses in WM and how people maintain the sequentially acquired data in WM. Nevertheless, a sequential model of hypothesis generation would allow one to model how different orders of the same data influence which hypotheses are generated and the order in which they are generated (Sprenger, 2007).

Cognitive–behavioral interventions. In addition to providing a useful metaphor for understanding how professional clinicians diagnose patients, HyGene also provides a theoretical framework for cognitive–behavioral therapy (CBT). CBT was founded, in part, on the idea that cognitions cause emotions (Beck, Rush, Shaw, & Emery, 1979). Beck and others have employed the construct of schemata as being central to CBT (see Brewin, 1996). According to Beck et al. (1979), many emotional problems can be traced to faulty or inaccurate schemata about the self or the external world. Thus, the route to treating emotional disorders is to treat the cognitions directly, by modifying the schema one entertains or by changing its content.

The construct of schema as employed in CBT parallels HyGene's semantic representations of hypotheses. Within HyGene, the consideration of faulty or inaccurate hypotheses can lead to biases in information processing. However, more generally, one can think of biases in information processing as arising from three possible sources: biases in the ecology (or one's representation of the ecology), biases in which memory cues (i.e., D_{obs}) are used to generate hypotheses, and/or inadequate generation of alternative hypotheses. The ecology in which one interacts defines the relationships between cues and hypotheses maintained in exemplar memory. Because exemplar memory is assumed to maintain the statistical relationships between cues and hypotheses, any biases in one's representation of the cue–hypothesis relationships can bias hypothesis generation. Biases in representation can arise either from true statistical relationships between D_{obs} and hypotheses present in the environment or through the failure to consider alternative hypotheses that might be associated with D_{obs}. For example, the overexposure to threat stimuli in one ecology (e.g., soldiers serving in a war zone) may lead to maladaptive responses in new ecologies (e.g., veterans who are stateside). Whereas it would be appropriate for a soldier in active combat to generate an ambush hypothesis in response to a loud boom, it would be maladaptive for a stateside veteran to generate the ambush hypothesis. Thus, HyGene places a premium on debiasing maladaptive prepotent responses to stimuli by restructuring statistical relationships maintained in episodic memory. This component of HyGene can be seen as having direct connections to behavioral therapies such as eye movement desensitization therapy (Shapiro, 1989), which aims to change maladaptive prepotent responses to environmental cues.

Alternatively, biases in the representation of cue–hypothesis relationships might arise from the failure to consider alternative hypotheses. It is well known in the judgment and decision-making literature that people often process data pseudodiagnostically (Doherty, Mynatt, Tweney, & Schiavo, 1979): They evaluate data only under one hypothesis, without examining how they relate to alternatives. Pseudodiagnostic search can lead to self-perpetuating biases in episodic memory. For example, people with depression who entertain the hypothesis "my friends do not like me" might be biased toward interpreting the behaviors of their friends as being supportive of their belief because they fail to consider whether the behavior is consistent with an alternative hypothesis. Such confirmation bias can in turn lead to biases in the statistical relationship between cues and hypotheses in episodic memory. Thus, interventions aimed at enticing patients to consider alternatives can lead to

a less biased interpretation of data as well as helping to correct (or prevent) biases in episodic memory. Note that the explicit consideration of hypotheses and the impact of these hypotheses on information search and judgment have obvious connections to cognitive therapy techniques. Indeed, many cognitive therapy techniques focus on improving patients' problem-solving skills and their ability to generate alternative solutions or explanations (Beck et al., 1979; D'Zurilla, & Goldfried, 1971). However, because hypothesis generation is assumed to be predicated on the statistical structures represented in exemplar memory, biases in one's representation of the ecology can lead to biases in hypothesis generation. Thus, HyGene anticipates that cognitive therapy will be most effective when accompanied by therapy aimed at correcting potential biases in one's representation of the ecology.

Debiasing judgment. As alluded to in the introduction, the findings within the hypothesis-generation, probability judgment, and hypothesis-testing literatures traditionally have been treated separately from one another. However, as should be obvious by now, substantial evidence suggests that these three areas are highly interrelated. Our goal in developing HyGene has been to illustrate the importance of hypothesis-generation processes for probability judgment.

Indeed, we have shown that many of the errors in the probability judgment literature are actually due to the input into the probability judgment process (the hypotheses that are generated) and not necessarily to biases in the computation of the probabilities. This finding is consistent with Tversky and Koehler's (1994) review indicating that the degree to which the set of alternative hypotheses is made explicit is related to judgment accuracy. Dougherty et al. (1997) showed that people who considered multiple alternative scenarios in a scenario-generation task gave lower confidence judgments to the most plausible causal scenario (i.e., suffered from less hindsight bias) than people who considered only one scenario. Dougherty and Hunter (2003a) revealed that the degree to which a focal hypothesis is overestimated is related to the number and strength of the alternative hypotheses considered by the decision maker: People who generate more alternative hypotheses provide lower and more accurate probability judgments (see also Pennington & Hastie, 1988).

Thus, in our view many of the ills of judgment arise due to a judge's failure to consider alternative hypotheses, which is, in part, a factor under the control of the decision maker. From the hypothesis-generation perspective, confidence judgments should become more accurate if one is encouraged to consider alternatives to the focal hypothesis (Koriat, Lichtenstein, & Fischhoff, 1980). Although there is some evidence that instructions to consider the opposite are effective at reducing overconfidence (Fischhoff & Downs, 1997; Koriat et al., 1980; Lord, Lepper, & Preston, 1984), such strategies are often ineffective. The framework proposed here suggests that instructions to consider the alternative hypotheses will be effective only when the alternatives are well specified. Our approach suggests that one should prompt decision makers to generate specific alternative hypotheses to the focal if such debiasing techniques are to be effective.

Extensions of the Model

Simultaneous versus serial presentation of cues. The simulations presented in this article have assumed that all the relevant cues are available to the decision maker and that all the cues are used simultaneously to probe episodic memory. However, in many real-world situations (e.g., the clinical and medical diagnostic situations), cues reveal themselves sequentially over time.

A variety of research supports the idea that sequential cue presentation affects judgment. For instance, Weber et al. (1993) found that the presentation order of symptoms resulted in different disease hypotheses being generated by clinicians. Also, the cue primacy effect suggests that decision makers weight or utilize the initial set of cues to a greater extent than later cues when generating and evaluating hypotheses (Adelman, Bresnick, Black, Marvin, & Sak, 1996). Extending the model to sequential cue presentation would allow it to handle tasks in which a cue's status changes over time. That is, HyGene could be extended to model hypothesis generation, evaluation, and testing in dynamic decision-making environments.

Discovering emerging hypotheses. Thus far, we have not discussed how HyGene might model cases in which the true or correct hypothesis is not in the decision maker's semantic memory. Indeed, as specified at the outset of our article, we treat hypothesis generation as a special case of cued recall, which presumably requires that the decision maker have knowledge of the possible hypotheses in semantic memory. However, how might one model generation processes when the decision maker encounters a unique pattern of cues in the environment that cannot be accounted for by hypotheses stored in semantic memory? In HyGene, novel hypotheses not contained within semantic memory are treated as aberrant cases, which are discovered as the result of the unspecified probe failing to match known hypotheses in semantic memory. This discovery process enables HyGene to generate novel hypotheses, use novel hypotheses to organize search, evaluate the efficacy of the novel hypotheses, and learn (i.e., remember novel hypotheses for future reference).

In medical diagnosis, the discovery process would be analogous to identifying a new syndrome, such as severe acute respiratory syndrome, that is characterized by a novel configuration of symptoms. The discovery of new hypotheses occurs when, for example, the unspecified probe does not match known hypotheses stored in semantic memory or when all of the generated hypotheses are rejected by the consistency-checking mechanism (i.e., the hypotheses in the SOC are eliminated because they are inconsistent with the presenting symptoms). When no hypotheses can be generated from semantic memory, the unspecified probe is assumed to be added to semantic memory as a unique (or aberrant) diagnostic hypothesis. Thus, the unspecified probe can be regenerated by future cases if it fits the data better than other hypotheses stored in semantic memory. Because episodic memory is updated on the basis of direct experience with the new hypothesis, the decision maker also has an exemplar of the new hypothesis and data stored in episodic memory. Moreover, the match between the unspecified probe and the semantic representation of that hypothesis will increase as the number of cases in one's experiences increases, which in turn should increase the probability that the newly discovered hypothesis is generated.

Although the HyGene discovery process has several practical implications, considerable empirical research is needed to understand how decision makers generate novel hypotheses. For instance, how dissimilar do aberrant cases have to be to what one already knows to engage the discovery process? Also, HyGene

postulates that decision makers use novel hypotheses to organize search and to evaluate the probability of novel hypotheses. The extent to which decision makers utilize novel hypotheses to accomplish creative discovery is neither well understood nor studied from a memory theoretic perspective.

HyGene as a General Model of Human Judgment

Our goal in this article has been to introduce a general theoretical framework that describes how participants generate hypotheses from long-term memory when presented with observable data and how those hypotheses influence both probability judgment and information search. Although several models have been developed to describe probability judgment and hypothesis-testing processes, no theory to date has been developed to describe hypothesis-generation processes, nor has there been any attempt to integrate hypothesis-generation processes with probability judgment and hypothesis-testing processes. In this sense, HyGene stands alone in its attempt to integrate these three literatures. Moreover, because HyGene is based on the MINERVA-DM model and incorporates support theory's comparison process, it can account for many (if not all) of the phenomena accounted for by MINERVA-DM (see Dougherty et al., 1999) and support theory (see Tversky & Koehler, 1994).

Although we view HyGene as a general model of judgment, it is an entirely content-based model. Whereas considerable evidence supports the idea that judgments are based on the contents of memory (Dougherty et al., 1999; Juslin & Persson, 2002; Reyna, 1991; Reyna & Brainerd, 1991; Sieck & Yates, 2001; Tversky & Kahneman, 1973), there also is evidence that people use information gleaned from metacognitive processes. For example, Schwarz et al. (1991; see also Sanna & Schwarz, 2004, 2006) proposed that judgments can be based on two distinct sources of information: the content of what is retrieved from memory and the subjective experience of the ease (or difficulty) of the retrieval process.

The phenomenological ease or difficulty of a retrieval task has been shown to influence judgment independent of the retrieved content (Sanna & Schwarz, 2004, 2006; Sanna, Schwarz, & Stocker, 2002; Schwarz et al., 1991). For example, in one study, Schwarz et al. (1991) found that participants who generated 12 instances of being assertive rated themselves as less assertive than participants who were required to generate only 3 instances of being assertive. Schwarz et al. argued that participants used the perceived difficulty of the retrieval task, rather than content, as the basis of their assertiveness judgments. Similar results have been found in a variety of judgment tasks (see Sanna & Schwarz, 2004). Because HyGene is formulated as a content-based model, it cannot account for the influence of metacognitive processes, such as the experienced ease or difficulty of retrieval, on judgment.

Our model is neither normative nor entirely semantic based but rather is a compromise between a Bayesian-like inference process and a semantic memory system. In this way, our model satisfies the definition of bounded rationality in that it utilizes algorithms borrowed from normative theory yet capitalizes on the informational value inherent in semantics and is constrained by the cognitive system. Our analysis is aimed at describing the processes underlying inductive inference in tasks where the data are probabilistically related to one or more potential hypotheses. While this type of task is characteristic of a large number of real-world inference tasks, it does not map well onto abductive inference tasks in which participants engage in chains of logical reasoning or where hypotheses are derived through analogy (Josephson & Josephson, 1994). Abductive inferences of these sorts are outside the scope of our model.

However, HyGene does have natural connections to exemplar-similarity models of categorization (e.g., Medin & Schaffer, 1978; Nosofsky, 1986; for a review, see Ashby & Maddox, 2005; for a formal comparison of likelihood and similarity-based models, see Nosofsky, 1990).[9] Many models of categorization are formulated as similarity-based models, where the task is to identify to which of N (where typically $N = 2$) categories a stimulus belongs. To be sure, HyGene can be considered a model of categorization, but one where the model generates the set of N alternative categories to which the stimulus might belong. Like models of categorization, HyGene can handle rapid or speeded perceptual categorization, as in the case where the data are uniquely related to a single hypothesis in semantic memory. This type of rapid categorization process is a logical candidate for describing what Klein (1993) called recognition-primed decision. However, HyGene can also handle more deliberative categorization, where the data are shared among a relatively large set of hypotheses in semantic memory and where the decision maker must choose one hypothesis (or category label) among a set of explicitly considered possibilities by comparing their relative probabilities—a process that is assumed to be capacity limited. Moreover, the model anticipates that an important component of deliberative categorization will be information search or hypothesis testing: Cases in which the data do not allow for a fast and confident determination of the correct hypothesis may require the decision maker to search for new information to discriminate among the set of potential hypotheses. Deliberative categorization processes of these sorts characterize expert physicians' generation and use of differential hypotheses in many diagnostic tasks (Elstein & Schwarz, 2002). In sum, HyGene can be considered a general model of judgment that describes both the processes of categorization (as well as category, or hypothesis, generation) and how the processes involved in categorization can affect probability judgment and information search. Thus, HyGene anticipates that there is a natural connection between the processes that underlie categorization processes and the processes that underlie judgment and decision making (Ashby & Berretty, 1997).

Conclusions

In this article, we have presented a new model of diagnostic hypothesis generation and human judgment. HyGene not only provides a cognitive process account of human judgment but goes further by modeling the relationship between the mental representation of an ecology, hypothesis generation, and judgment. HyGene anticipates that which hypotheses one generates will be constrained by one's mental representation of the ecology. Moreover, HyGene predicts that errors or biases that creep into the hypothesis-generation process will cascade into errors and biases

[9] The precursor to HyGene, MINERVA-DM, can be considered an exemplar-similarity model of categorization, where categorization is based on the Bayesian probability that object i belongs to category A given a pattern of features (see the Appendix and Dougherty et al., 1999).

in probability judgment and information search. Importantly, HyGene's account of phenomena within the judgment and decision-making literature does not rely on the ad hoc application of heuristic mechanisms. Instead, HyGene provides an integrative theoretical framework linking the processes of hypothesis generation and probability judgment to the underlying processes of memory. Although we espouse the view held by ecological psychologists who place a premium on modeling behavior as a function of the ecology (e.g., Gigerenzer et al., 1991), we suggest that the influence of the ecology is realized through basic memory processes. In this way, HyGene departs from prior approaches to judgment and decision making that view judgment processes as consisting of a collection of semiautonomous domain-specific heuristic mechanisms (Gigerenzer & Goldstein, 1996; Gigerenzer, Todd, & the ABC Group, 1999). Although people likely employ heuristic mechanisms in judgment and choice, we argue that the memory system places constraints on the operation of these heuristics. Rather than assuming that the mind has evolved a variety of special purpose heuristics, our approach has been to model judgment and decision making as a function memory, having assumed that memory processes have adapted to serve multiple functions, including judgment and decision making.

References

Adelman, L., Bresnick, T., Black, P. K., Marvin, F. F., & Sak, S. G. (1996). Research with Patriot air defense officers: Examining information order effects. *Human Factors, 38,* 250–261.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–419.

Arnott, D., & Pervan, G. (2005). A critical analysis of decision support systems research. *Journal of Information Technology, 20,* 67–87.

Ashby, F. G., & Berretty, P. M. (1997). Categorization as a special case of decision-making or choice. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (pp. 367–388). Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149–178.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8, pp. 47–90). New York: Academic Press.

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

Bearden, J. N., & Wallsten, T. S. (2004). MINERVA-DM and subadditive frequency judgments. *Journal of Behavioral Decision Making, 17,* 349–363.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy for depression.* New York: Wiley.

Botti, M., & Reeve, R. (2003). Role of knowledge and ability in student nurses' clinical decision making. *Nursing and Health Sciences, 5,* 39–49.

Brewin, C. R. (1996). Theoretical foundations of cognitive-behavior therapy for anxiety and depression. *Annual Review of Psychology, 47,* 33–57.

Caillies, S., Denhiere, G., & Kintsch. W. (2002). The effect of prior knowledge on understanding from text: Evidence from primed recognition. *European Journal of Cognitive Psychology, 14,* 267–286.

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 and Brain Sciences, 24,* 87–185.

D'Zurilla, T., & Goldfried, M. (1971). Problem solving and behavior modification. *Journal of Abnormal Psychology, 78,* 107–126.

Doherty, M. E., Chadwick, R., Garavan, H., Barr, D., & Mynatt, C. R. (1996). On people's understanding of the diagnostic implications of probabilistic data. *Memory & Cognition, 24,* 644–654.

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

Dougherty, M. R., & Harbison, J. I. (2007). Motivated to retrieve: How often are you willing to go back to the well when the well is dry? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 1108–1117.

Dougherty, M. R., & Sprenger, A. (2006). The influence of improper sets of information on judgment: How irrelevant information can bias judged probability. *Journal of Experimental Psychology: General, 135,* 262–281.

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

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

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

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

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

Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *British Medical Journal, 324,* 729–732.

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

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.

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 349–358.

Fischhoff, B., & Downs, J. (1997). Accentuate the relevant. *Psychological Science, 8,* 1–5.

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

Flin, R., Slaven, G., & Stewart, K. (1996). Emergency decision making in the offshore oil and gas industry. *Human Factors, 38,* 262–277.

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

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

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

Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.

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

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Gordon, B. L. (1970). Terminology and content of the medical record. *Computational Biomedical Research, 3,* 436–444.

Hájek, A. (2007). The reference class problem is your problem too. *Synthese, 156,* 563–585.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior, 16,* 321–338.

Hintzman, D. L. (1984). MINERVA2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers, 16,* 96–101.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Hintzman, D. L. (1987). Recognition and recall in MINERVA 2: Analysis of the "recognition failure" paradigm. In P. Morris (Ed.), *Modeling cognition* (pp. 215–229). London: Wiley.

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

Joseph, G. M., & Patel, V. L. (1990). Domain knowledge and hypothesis generation in diagnostic reasoning. *Medical Decision Making, 10,* 31–46.

Josephson, J. R., & Josephson, S. J. (1994). *Abductive inference: Computation, philosophy, technology.* New York: Cambridge University Press.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26,* 563–607.

Kilinç, B. E. (2001). The reception of John Venn's philosophy of probability. In V. F. Hendricks, S. A. Pedersen, & K. F. Jorgensen (Eds.), *Probability theory: Philosophy, recent history, and relations to science* (pp. 97–124). Dordrecht, the Netherlands: Kluwer Academic.

Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.), *Thematics: Interdisciplinary studies* (pp. 157–170). Amsterdam: John Benjamins.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363–394.

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

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In J. Orasanu (Ed.), *Decision making in action: Models and methods* (pp. 138–147). Westport, CT: Ablex Publishing.

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

Koehler, D. J., Brenner, L. A., Liberman, V., & Tversky, A. (1996). Confidence and accuracy in trait inference: Judgment of similarity. *Acta Psychologica, 92,* 33–57.

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

Kruglanski, A. W., & Mayseless, O. (1988). Contextual effect in hypothesis testing: The role of competing alternatives and epistemic motivations. *Social Cognition, 6,* 1–20.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing." *Psychological Review, 103,* 263–283.

Leblanc, V. R. (2003). The influence of a tentative diagnosis on the identification of features from patient appearance. *Dissertation Abstracts International, 63,* 4375B.

Libby, R. (1985). Availability and the generation of hypotheses in analytical review. *Journal of Accounting Research, 23,* 646–665.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47,* 1231–1243.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

Mayes, A. R., & Montaldi, D. (2001). Exploring the neural bases of episodic and semantic memory: The role of structural and functional neuroimaging. *Neuroscience & Biobehavioral Reviews, 25,* 555–573.

Mayseless, O., & Kruglanski, A. W. (1987). What makes you so sure? Effects of epistemic motivations on judgmental confidence. *Organizational Behavior and Human Decision Processes, 39,* 162–183.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

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.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology, 34,* 393–418.

Park, D. C., Smith, A. D., Lautenschlager, G., & Earles, J. L. (1996). Mediators of long-term memory performances across the life span. *Psychology and Aging, 11,* 621–637.

Patrick, J., Grainger, L., Gregov, A., Halliday, P., James, N., & O'Reilly, S. (1999). Training to break the barriers of habit in reasoning about unusual faults. *Journal of Experimental Psychology: Applied, 5,* 314–355.

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: The effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 521–533.

Rawson, K. A., & Kintsch, W. (2002). How does background information improve memory for text content*? Memory & Cognition, 30,* 768–778.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382–407.

Reyna, V. F. (1991). Class inclusion, the conjunction fallacy, and other cognitive illusions. *Developmental Review, 11,* 317–336.

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making, 4,* 249–262.

Sanna, L. J., & Schwarz, N. (2004). Integrating temporal biases: The interplay of focal thoughts and accessibility experiences. *Psychological Science, 15,* 474–481.

Sanna, L. J., & Schwarz, N. (2006). Metacognitive experiences and human judgment: The case of hindsight bias and its debiasing. *Current Directions in Psychological Science, 15,* 172–176.

Sanna, L. J., Schwarz, N., & Stocker, S. L. (2002). When debiasing backfires: Accessible content and accessibility experiences in debiasing hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 497–502.

Schwarz, N., Bless, H., Strack, F., Klump, G. Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61,* 195–202.

Shapiro, F. (1989). Efficacy of the eye movement desensitization procedure in the treatment of traumatic memories. *Journal of Traumatic Stress, 2,* 199–223.

Shim, J. P., Warkenten, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems, 33,* 111–126.

Sieck, W. R. & Yates, J. F. (2001). Overconfidence effects in category learning: A comparison of connectionist and exemplar memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1003–1021.

Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence, 4,* 181–201.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived confirmation. *Journal of Experimental Social Psychology, 22,* 93–121.

Slowiaczek, L. M., Klayman, J., & Sherman, S. J. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition, 20,* 392–405.

Sprenger, A., & Dougherty, M. R. (2006). Differences between probability and frequency judgments: The role of individual differences in working memory capacity. *Organizational Behavior and Human Decision Processes, 99,* 202–211.

Sprenger, A. M. L. (2007). *Sequential hypothesis generation.* Doctoral dissertation, University of Maryland, College Park. Retrieved November 25, 2007, from ProQuest Digital Dissertations database (Publication No. AAT 3260299).

Stanovich, K. E., & West, R. F. (1998a). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127,* 161–188.

Stanovich, K. E., & West, R. F. (1998b). Who uses base rates and P(D|−H)? An analysis of individual differences. *Memory & Cognition, 28,* 161–179.

Stausberg, J., & Person, M. (1999). A process model of diagnostic reasoning in medicine. *International Journal of Medical Informatics, 54,* 9–23.

Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology, 43,* 22–34.

Trope, Y., & Mackie, D. M. (1987). Sensitivity to alternatives in social hypothesis-testing. *Journal of Experimental Social Psychology, 23,* 445–459.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53,* 1–25.

Tulving, E., Hayman, C. A., & Macdonald, C. A. (1991). Long-lasting perceptual priming and semantic learning in amnesia: A case experiment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 595–617.

Tulving, E., & Markowitsch, H. J. (1998). Episodic and declarative memory: Role of the hippocampus. *Hippocampus, 8,* 198–204.

Tulving, E., Schacter, D. L., McLachlan, D. R., & Moscovitch, M. (1988). Priming of semantic autobiographical knowledge: A case study of retrograde amnesia. *Brain and Cognition, 8,* 3–20.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 201–232.

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

Vermande, M. M., van den Bercken, J. H., & De Bruyn, E. E. (1996). Effects of diagnostic classification systems on clinical hypothesis-generation. *Journal of Psychopathology and Behavioral Assessment, 18,* 49–70.

Ward, J. (2003). Encoding fidelity and the frontal lobes: A dissociation between retrograde and anterograde memories. *Cortex, 39,* 791–812.

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

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67,* 1049–1062.

Webster, D. M., Richter, L. & Kruglanski, A. W. (1996). On leaping to conclusions when feeling tired: Mental fatigue effects on impressional primacy. *Journal of Experimental Social Psychology, 32,* 181–195.

Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 198–215.

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

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

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

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

Appendix

HyGene Calculation Example

Below, we illustrate the computational details of HyGene in a simplified ecology with three events. The three events share 75% of their features with two 9-element random minivectors. Although we use 15-element minivectors in our simulations, we use 9-element vectors for the example to save space. The calculation example follows the steps illustrated in Figure 2 in the main text.

The number of traces of each event in episodic memory is given by the event frequency in Table A1. The traces are encoded into episodic memory with encoding fidelity L = .85. In Step 1, the D_{obs} (the grey shaded vector illustrated in the Figure A1) is sampled from or observed in the environment. The data act as a retrieval probe that initiates the activation (A_i) of traces in episodic memory. In the example, the D_{obs} is the data component of Event 1.

The calculation example below (see Figure A2) illustrates how trace activation is derived using Trace 1 as an example. To calculate the activation of Trace 1 in response to D_{obs}, we take the cubed dot product between the probe vector and the data component (data minivector) for Trace 1, where P_j is a feature in the *j*th position of the probe, T_{1j} is a feature in the *j*th position of Trace 1, and N_i = number of features where $P_j \neq 0$ or $T_{1j} \neq 0$. There are four features that have a dot product of 1 between Trace 1 and the D_{obs}, $P_jT_{1j} = 4$. There are six features where $P_j \neq 0$ or $T_{1j} \neq 0$, so $N_1 = 6$. Thus, the activation of Trace 1 in response to the D_{obs} is

$$\left(\frac{4}{6}\right)^3 = .296$$

In Step 2 (see Figure A3), the traces activated above threshold value ($A_c = .216$ for the example) results in the extraction, from episodic memory, of an unspecified probe that resembles those hypotheses that are most commonly (and strongly) associated with the D_{obs}. We refer to the process of extracting an unspecified probe as conditional echo content with resolution. First, we assume that the conditional echo content vector is based only on those traces that are activated above threshold, $A_c = .216$, by the D_{obs} (i.e., data component of Event 1 in the example).

The value of A_i used to compute the content vector is based only on the activation of the data component of the probe (D_{obs}) and the corresponding (data) component of the trace. One can think of the activation of the data component of a trace being passed to the other components of the trace. To compute conditional echo content for each element of the hypothesis component, one first multiplies each feature value of a trace by the activation of the corresponding data component of the trace and then sums the resulting values across traces. For example, the first conditional echo content vector element in the hypothesis component is

$$\mathbf{C_C} = \sum_{i=1}^{K} A_i T_{ij} = ((.2963 \times 1) + (1 \times 1) + (.5787 \times 1) + (1 \times 1) + (1 \times 1) + (.5787 \times 0) + (\text{N/A}) + (\text{N/A}) + (\text{N/A}) + (\text{N/A})) = 3.87.$$

This process is repeated for the other elements in the conditional echo content vector for the hypothesis component as well as for the nine elements in the data component. The component conditional echo content vectors are then concatenated, and the content values are normalized by dividing by the absolute value of the largest content value. This process is referred to as conditional echo content resolution and ensures that content values are perceived within the allowable feature range of 1 to −1, while preserving the

Table A1
Example Ecology Table

	Example Ecology		
Event	E1	E2	E3
Event frequency	6	3	1
Prior probability	0.6	0.3	0.1

	Data Component									
Probe (Dobs E1)	0	1	-1	1	0	1	-1	1	0	A_i
Trace1 (E1)	0	0	-1	1	0	1	-1	0	0	0.2963
Trace2 (E1)	0	1	-1	1	0	1	-1	1	0	1
Trace3 (E1)	0	1	-1	0	0	1	-1	1	0	0.5787
Trace4 (E1)	0	1	-1	1	0	1	-1	1	0	1
Trace5 (E1)	0	1	-1	1	0	1	-1	1	0	1
Trace6 (E1)	0	0	-1	1	0	1	-1	1	0	0.5787
Trace7 (E2)	-1	0	-1	0	-1	1	1	0	0	0.002
Trace8 (E2)	-1	0	-1	1	-1	1	1	0	0	0.0156
Trace9 (E2)	-1	0	0	1	-1	1	1	1	0	0.0156
Trace10 (E3)	1	0	-1	-1	1	1	-1	0	0	0.0156

Figure A1. Step1: Activation of traces in response to D_{obs}. A_i = activation of $Trace_i$; D_{obs} = pattern of observable data; E = event.

	Data Component								
Probe (Dobs E1)	0	1	-1	1	0	1	-1	1	0
TRACE (Trace1)	0	0	-1	1	0	1	-1	0	0
P_jT1_j	0	0	1	1	0	1	1	0	0
N_i	0	1	1	1	0	1	1	1	0

$$A_i = \left(\frac{\sum_{j=1}^{N} P_j T_{1j}}{N_1}\right)^3 = \left(\frac{4}{6}\right)^3 = .296$$

Figure A2. Activation calculation example. A_i = activation of $Trace_i$; D_{obs} = pattern of observable data; E = event; N_i = number of features where $P_j \neq 0$ or $T_{1j} \neq 0$; P_j = features in the *j*th position of the probe; T_{1j} = features in the *j*th position of Trace 1.

(*Appendix continues*)

	Data Component									A_i	Pass A_c	*Hypothesis Component*								
Probe (Dobs E1)	0	1	-1	1	0	1	-1	1	0											
Trace1 (E1)	0	0	-1	1	0	1	-1	0	0	0.2963	Yes	1	-1	0	0	1	1	-1	1	0
Trace2 (E1)	0	1	-1	1	0	1	-1	1	0	1	Yes	1	-1	0	0	1	1	-1	1	0
Trace3 (E1)	0	1	-1	0	0	1	-1	1	0	0.5787	Yes	1	-1	0	0	1	1	-1	1	0
Trace4 (E1)	0	1	-1	1	0	1	-1	1	0	1	Yes	1	0	0	0	1	1	-1	1	0
Trace5 (E1)	0	1	-1	1	0	1	-1	1	0	1	Yes	1	0	0	0	1	1	-1	1	0
Trace6 (E1)	0	0	-1	1	0	1	-1	1	0	0.5787	Yes	0	0	0	0	1	1	-1	1	0
Trace7 (E2)	-1	0	-1	0	-1	1	1	0	0	0.002	No	1	-1	0	0	0	1	0	1	0
Trace8 (E2)	-1	0	-1	1	-1	1	1	0	0	0.0156	No	0	-1	0	0	0	1	0	1	0
Trace9 (E2)	-1	0	0	1	-1	1	1	1	0	0.0156	No	1	-1	0	0	0	1	0	1	0
Trace10 (E3)	-1	0	-1	-1	1	1	-1	0	0	0.0156	No	1	0	0	0	0	0	-1	1	0
	0	3.58	-4.5	3.87	0	4.45	-4.5	4.16	0			3.87	-1.9	0	0	4.45	4.45	-4.5	4.45	0
Unspecified Probe	0	0.8	-1	0.87	0	1	-1	0.93	0	0.87	-0.4	0	0	1	1	-1	1	0		

Figure A3. Step 2: Creation of an unspecified probe (HyGene's conditional echo content where $A_c = .216$). A_c = activation criterion; A_i = activation of Trace_i; D_{obs} = pattern of observable data; E = event.

	Data Component									*Hypothesis Component*									A_s
Unspecified Probe	0	1	-1	1	0	1	-1	1	0	1	-0	0	0	1	1	-1	1	0	
Semantic Trace (E-1)	0	1	-1	1	0	1	-1	1	0	1	-1	0	0	1	1	-1	1	0	0.7491
Semantic Trace (E-2)	-1	0	-1	1	-1	1	1	1	0	1	-1	0	0	0	1	0	1	0	0.0825
Semantic Trace (E-3)	1	0	-1	-1	1	1	-1	0	0	1	0	0	0	0	1	-1	1	0	0.0787

Figure A4. Step 3: Semantic activation to unspecified probe. A_s = activation of semantic Trace_s; E = event.

sign of the original content values. The data component of the unspecified probe always closely resembles the D_{obs} that produced it, and the hypothesis component resembles the hypothesis (or hypotheses) most often and most strongly associated with the D_{obs} for the traces in memory whose activation exceeds threshold.

In Step 3 (see Figure A4), the unspecified probe serves as the basis for the process of hypothesis generation. The determination of what the unspecified probe might represent is achieved by matching it against semantic memory, where the semantic traces are assumed to represent known hypotheses. For simplicity, we assume that the unspecified probe activates all semantic traces in parallel. Hypotheses in semantic memory with activation (A_s) greater than zero are assumed to define the set of relevant hypotheses from which the decision maker samples when generating hypotheses.

The probability that a semantic trace is sampled for generation is given by its activation relative to the activation of all other traces in semantic memory with positive activation (i.e., normed A_s). For instance (see Figure A5), the normed A_s for Semantic Trace 1 is .82, which means that for every generation attempt, there is an 82% chance that Semantic Trace 1 will be sampled. Sampled traces are then compared with the threshold for generation, which we refer to as Act_{MinH}. Act_{MinH} has an initial value of zero but is dynamically updated based on the activation values of semantic traces that have been passed into the set of leading contenders (SOC). Failure of a semantic trace to pass Act_{MinH} is considered a retrieval failure. Moreover, the sampling from semantic memory of a trace that already resides in the SOC is also counted as a retrieval failure. Hypothesis generation stops when the successive number of retrieval failures equals TMAX.

	Normed A_s
Semantic Trace (E-1)	0.8229
Semantic Trace (E-2)	0.0906
Semantic Trace (E-3)	0.0864

Figure A5. Step 4: Probability that a semantic trace is compared with Act_{MinH} (normed A_i). A_i = activation of Trace_i; A_s = activation of semantic Trace_s; Act_{MinH} = minimum activation threshold for hypothesis to enter the set of leading contenders; E = event.

In the Step 5 calculation example (see Figure A6), we illustrate how the memory strength of a particular leading contender is derived. This process is used as input into HyGene's comparison process to produce posterior probabilities. In the calculation example, we assume that H_1, which is the hypothesis component of Semantic Trace 1, has been successfully generated into the SOC and is a leading contender. The memory strength of a leading contender is conditional on D_{obs}. Only traces that are activated above threshold ($A_c = .216$ for the example) contribute to the memory strength (activation) of H_1, which is referred to as conditional echo intensity. In the example, the data component of the probe (i.e., D_{obs}) is the data component of Event 1. The echo intensity of H_1 is given by

$$I_C = \frac{\sum I_{A_i \geq A_c}}{K}$$

the sum of the activations across the K traces in the activated subset, where I_c is the mean echo intensity (4.45 in the example) and K is the number of traces for which $A_i > A_c$ (six in the example).

We assume that posterior probability judgments result from a comparison process that operates on the conditional echo intensities of the leading contender hypotheses (i.e., the SOC). The judged posterior probability of a hypothesis is given by

$$P(H_i|D_{obs}) = \frac{I_{C_i}}{\sum_{i=1}^{W} I_{C_i}}$$

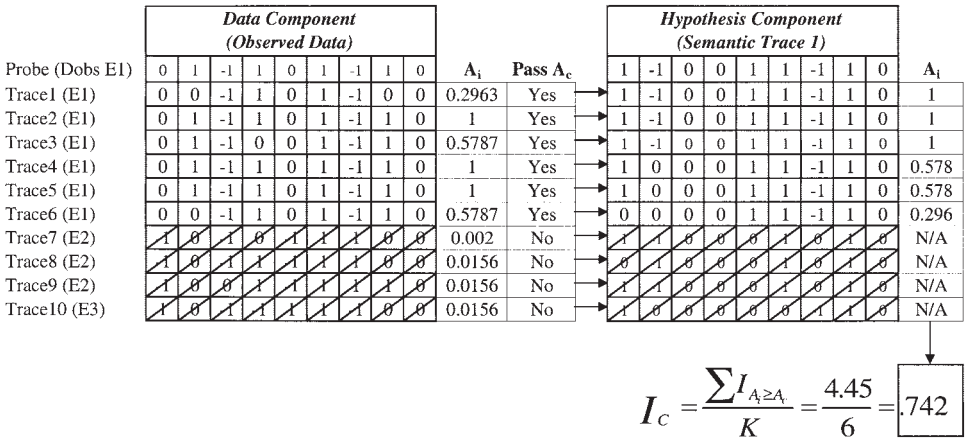

Figure A6. Step 5: Conditional echo intensity for $H_1|D_{obs}$. A_c = activation criterion; A_i = activation of $Trace_i$; D_{obs} = pattern of observable data; E = event; H_1 = Hypothesis 1; I = echo intensity; I_c = conditional echo intensity; K = number of traces for which $A_i > A_c$.

Scenario 1	**SOC**	*Conditional Echo Intensity*	*Sum Echos*	*Posterior Probability Judgment*
	H1 (Retrieval Failure)	0	0	0
Scenario 2	**SOC**			
	H1	0.742	0.742	1
Scenario 3	**SOC**			
	H1	0.742	0.938	0.791044776
	H2	0.196		0.208955224
Scenario 4	**SOC**			
	H1	0.742	1.009	0.735381566
	H3	0.267		0.264618434
Scenario 5	**SOC**			
	H1	0.742	1.205	0.615767635
	H2	0.196		0.162655602
	H3	0.267		0.221576763

Figure A7. Step 6: Posterior probability judgment scenarios for H_1. H_1 = Hypothesis 1; SOC = set of leading contenders.

or the mean conditional echo intensity of the leading contender normalized by the sum of the conditional echo intensities of all leading contenders in the SOC. Where $P(H_i|D_{obs})$ is the probability of the *i*th hypothesis in the SOC, conditional on the subset of traces by D_{obs} (the data observed in the environment that was initially used to partition the subset of activated traces in episodic memory). In the example Step 6 (see Figure A7), we derive every possible configuration of the SOC that includes H_1 to illustrate how HyGene derives posterior probability judgments. In the scenarios below, the model is always asked to judge the posterior probability of H_1 in response to the D_{obs}. In Scenario 1, the model fails to activate any traces above threshold. Under these conditions, we assume the posterior judgment of H_1 is 0. In Scenario 2, either H_1 is the only hypothesis generated by the model or no hypotheses are generated and H_1 is prompted by the elicitation. In either case, H_1 is the only hypothesis in the SOC, and its posterior probability is judged 1. In Scenario 3, either or both H_1 and H_2 are generated by the model in response to D_{obs} or H_2 is generated by the model in response to D_{obs} while H_1 is prompted by the probability elicitation. Note that the judged posterior probability of H_1 is less when the comparison process includes another hypothesis (H_2 in Scenario 3, H_3 in Scenario 4, and both H_2 and H_3 in Scenario 5). Thus, the fewer leading contender alternatives there are to H_1, the higher H_1's subjective posterior probability. This property of the model is the basis of its predictions of excessive posterior probability judgments (i.e., subadditivity). The posterior probability judgments show additivity over the proper set only when all proper set hypotheses are leading contenders (i.e., Scenario 5, where all three hypotheses are generated). However, the sum of the posterior probability judgments of the leading contenders is 1 (constrained additivity) for any particular configuration of the SOC. Thus, subadditivity can be considered a consequence of HyGene's constrained additivity property.

Received September 13, 2005
Revision received August 28, 2007
Accepted September 5, 2007 ■