Research Article

Psychological Science
2022, Vol. 33(3) 450–462

DOI: 10.1177/09567976211043425
www.psychologicalscience.org/PS

Strength of Belief Guides Information Foraging

David A. Illingworth[1] and Rick P. Thomas[2]
[1]Department of Psychology, University of Maryland, College Park, and [2]School of Psychology, Georgia Institute of Technology

Abstract

Theories of how people value and search for information share the assumption that beliefs give rise to the perceived value of information. However, few studies have directly addressed the pre-search processes that influence information-foraging behavior. This experiment examined the influence of pre-search belief updating on the perceived value of information sources. A sample of college students completed a hypothesis-testing, medical-diagnosis task. The experiment used medical tests with equal objective informative value before unveiling a presenting symptom intended to alter the strength of belief in different disease hypotheses. The observed patterns of test selection suggest that changes in beliefs about disease hypotheses result in systematic and predictable changes in test preference—a notion we refer to as the principle of *hypothesis-guided search*. We also present a simulation of how pre-search processes (e.g., hypothesis generation and working memory capacity) and task variables (e.g., time pressure) influence subsequent information search.

Keywords

decision-making, hypothesis testing, information search, open data, open materials, preregistered

Received 9/2/20; Revision accepted 8/3/21

Many instances of human judgment occur in environments of high uncertainty, and decision-makers should consider multiple hypotheses as candidate explanations for a set of observations. Decision-makers often search for information under such circumstances to test hypotheses and improve their judgments and decisions. Consider, as an example, the diagnostic tasks carried out by physicians. Their information-search behavior—the selection or ordering of medical tests—is preceded by at least one initial observation of a patient or a summary of the case. The disease hypotheses considered as possible explanations of those initial observations, and the strength of these beliefs, are critical to physicians' valuation of information sources. Ignoring hypotheses misses the underlying motivation and context for their search—evaluating the likely causes of the patient's ailment.

The process by which people decide which information and which information sources are potentially useful is an important topic of study. Some researchers have focused on characterizing the perceived value of information (Evans & Over, 1996; Kirby, 1994; Klayman & Ha, 1987; Manktelow & Over, 1990; Oaksford & Chater, 1996; Poletiek, 1995), whereas others have examined the act of seeking and acquiring information (Cohen et al., 2007; Hills, 2006; Hills et al., 2012, 2015; Mehlhorn et al., 2015; Pirolli & Card, 1999). A human ability to detect the value of information, often without explicit or direct access, is at the core of these ideas.

The literature has generally accounted for people's sensitivity to the usefulness of information via the formation of (pre-search) expectations that gauge the value of potential answers to the queries that people formulate (Johnson-Laird & Byrne, 1991). The notion that expectations drive search has prompted many researchers to investigate information metrics that explain people's information-search behavior (Nelson, 2005; Nelson et al., 2010). What remains untested is the fundamental assumption that links the formation of

Corresponding Author:
David A. Illingworth, University of Maryland, College Park, Department of Psychology
Email: davidai@umd.edu

expectations (via belief updating) following early data acquisition to subsequent search and hypothesis-testing behaviors (Coenen et al., 2018).

Traditionally, a researcher might manipulate the base rates of hypotheses or the data (possible test outcomes) to change one's initial beliefs. In practice, the base rates of hypotheses (e.g., diseases) and data are stable; professionals are required to generate the relevant hypotheses themselves. We suggest that the products of the pre-search processes (hypothesis generation and belief updating) determine the initial strength of belief in the hypothesis set, making the pre-search mechanisms applicable to all theories accounting for how people value and search for information.

Prior research has reported links between pre-search belief updating and patterns of information preference in hypothesis-testing tasks. Dougherty et al. (2010) found evidence that pre-search belief-updating processes accounted for preferences for pseudodiagnostic and diagnostic tests. Specifically, individuals who generated only a single hypothesis were more likely to exhibit a preference for pseudodiagnostic tests, suggesting that confirmation bias can be accounted for by the generation or consideration of a single hypothesis. Participants who generated more than one hypothesis were more likely to exhibit a preference for more diagnostic sources of information. Similarly, Buttaccio et al. (2015, 2018) examined the role of hypothesis generation in a visual search task. They developed a retrieval-guidance paradigm in which possible target representations were generated from long-term memory on the basis of external cues. The eye-tracking data were best fitted by a model of hypothesis generation in which the first target characteristic retrieved from memory guided visual search.

The HyGene architecture (Thomas et al., 2008) predicts that the hypotheses under consideration by a decision maker, and the relative memory-driven support of each, influence test preferences beyond the mere number of hypotheses believed to be in contention. In the HyGene model, the strength of belief in a hypothesis is a function of both the support from memory activation and the number of competing hypotheses maintained in a set of contenders (SOC)—a working memory construct limited in capacity by both cognitive (e.g., individual differences) and task (e.g., time pressure) constraints. The SOC's hypotheses are available as input for additional tasks such as hypothesis testing. We treated hypothesis-guided search as a dynamic, retrieval-driven process that provides a formal theory of information search—the value or preference exhibited for information sources changes because of pre-search memory mechanisms and belief updating.

Figure 1 illustrates the predicted influence of hypothesis generation on information search via cued recall (memory retrieval) using HyGene. Again, the simulation's underlying assumption is that the set of generated hypotheses constrains information search while cued recall governs the composition and number of hypotheses in the SOC. How hypotheses are generated into the SOC is detailed elsewhere (Thomas et al., 2008) and in the Supplemental Material available online. We extended HyGene to model valuation judgments for sources of information (T) using a bounded rational form of Bayesian diagnosticity (Equation 1). For each observable outcome (d_j), a maximum likelihood ratio was computed using the memory activation in support of each outcome conditional on using each candidate hypothesis (SOC_i) relative to the same activation conditional on all competing hypotheses (SOC_0). The mean maximum likelihood across all candidate hypotheses was weighted by the total memory activation in support of the outcome. The value of each source of information was the sum of all weighted, maximum likelihoods:

$$\text{value}\left(T\right) = \sum_{j=1}^{d} \left(\text{Act}\left(d_j\right) \times \frac{\sum_{i=1}^{i} \max\left(\frac{\text{Act}(d_j) \mid \text{SOC}_i}{\text{Act}(d_j) \mid \text{SOC}_0}, \frac{\text{Act}(d_j) \mid \text{SOC}_0}{\text{Act}(d_j) \mid \text{SOC}_i} \right)}{\text{SOC}_{\text{length}}} \right) \quad (1)$$

Statement of Relevance

How people formulate questions and challenge their beliefs have long motivated research in psychological science. Many researchers have proposed that the statistical properties of the information sources that people use to answer questions or challenge beliefs account for how they decide which information to look for. We tested an alternative assumption made by virtually all theories of information search—that strength of belief determines perceptions of value in information sources. This research found evidence for a process we refer to as *hypothesis-guided search*, because the information sources that people preferred changed in response to an initial piece of information specifically designed to bias their strength of belief in candidate diseases in a medical-diagnosis task. This finding is important because it clearly illustrates the importance of the hypotheses people entertain as explanations for their observations when they decide to seek out additional information.

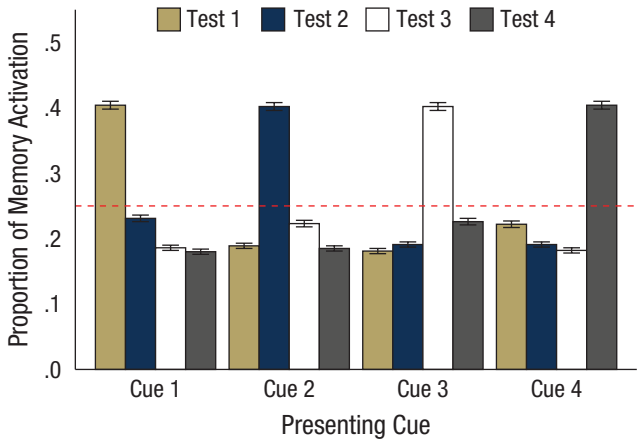

Fig. 1. Simulation data showing predicted memory activation under the HyGene model for each of four presenting cues and each of four tests. The standardized memory activations plotted in this figure are theorized to motivate the preference for each of the four available tests. Activation patterns changed because of changes in belief driven by observation of the presenting cue. The dashed line represents the standardized memory activation for all tests before any cue is presented. Error bars represent standard errors.

We crafted a data environment in which each of four cues possessed strong relations with two of four possible disease hypotheses, and four equally informative tests (before cue presentation) were available to further disambiguate the disease afflicting a simulated patient (see Table 1). The base rates for all hypotheses, presenting cues, and test outcomes were controlled to be equal. Thus, only hypothesis generation and belief-updating processes initiated by observation of the presenting cues could account for differential patterns of memory activation that indicated the model's preference for each available test. Observation of Cues 1, 2, 3, and 4 led to preferences for Tests 1, 2, 3, and 4, respectively (Fig. 1). The model simulated 200 decision-makers endowed with a memory of 800 patients.

We designed the data environment's statistical structure to impose some control over the hypotheses most strongly considered by the participants, which we argue is necessary to capture the dynamic nature of belief updating and information valuation. Our empirical hypothesis was that different presenting cues would cause markedly different test preferences consistent with the model's behavior. It is important to note that HyGene's predictions are nominally consistent with those of numerous normative models (see Fig. 2 for an illustration). However, HyGene provides a cognitive-process account for belief strength.

Method

This protocol was reviewed and approved by the institutional review board at the Georgia Institute of Technology. The preregistration for the study is available on OSF (https://osf.io/67z5v/).

Participants

Undergraduate students enrolled at the Georgia Institute of Technology participated in this study via an online experiment-management system (Sona Systems; https://sona-systems.com/). This population was selected to acquire a convenience sample. In total, 56 participants (22 women) completed the experiment. An additional nine participants were sampled but excluded from analysis because of computer error. We assumed a medium effect size and targeted a sample size of 60. All participants received partial course credit for their involvement in the study.

Design

Hypothesis-guided search assumes that expectations formed via belief updating and hypothesis generation drive information preference. Detecting such a process required the careful construction of cues that would direct participants to entertain specific sets of disease hypotheses and control the distribution of those beliefs (i.e., the strength of belief). The data environment that defined the relation between disease states and presenting cues is outlined in Table 1. The values in Table 1 represent the conditional probability that each of the four equally informative symptoms would manifest given that the patient was experiencing one of the four disease hypotheses.

This relation between presenting cues and disease hypotheses was the basis of the within-subjects design of this experiment. All participants learned this same ecological data structure through experience, leveraging this knowledge in a test phase in which trials would begin with access to the presenting cue only. Examination of the presenting-cue columns in Table 1 reveals the objective belief distribution across the disease states. Given perfect knowledge of the presenting-cue dynamic, presentation of a different presenting symptom results in drastically different belief states. We explain the protocol in greater detail in the Procedure section.

Table 2 lists the objective statistical structure for the medical tests available to the participants. The values

Table 1. Objective-Presenting-Symptom Statistical Environment

Hypothesis	Cue 1	Cue 2	Cue 3	Cue 4
Hypothesis 1	.50	.10	.10	.30
Hypothesis 2	.30	.50	.10	.10
Hypothesis 3	.10	.30	.50	.10
Hypothesis 4	.10	.10	.30	.50

Note: Values represent the probability of observing each presenting symptom given the underlying disease hypothesis.

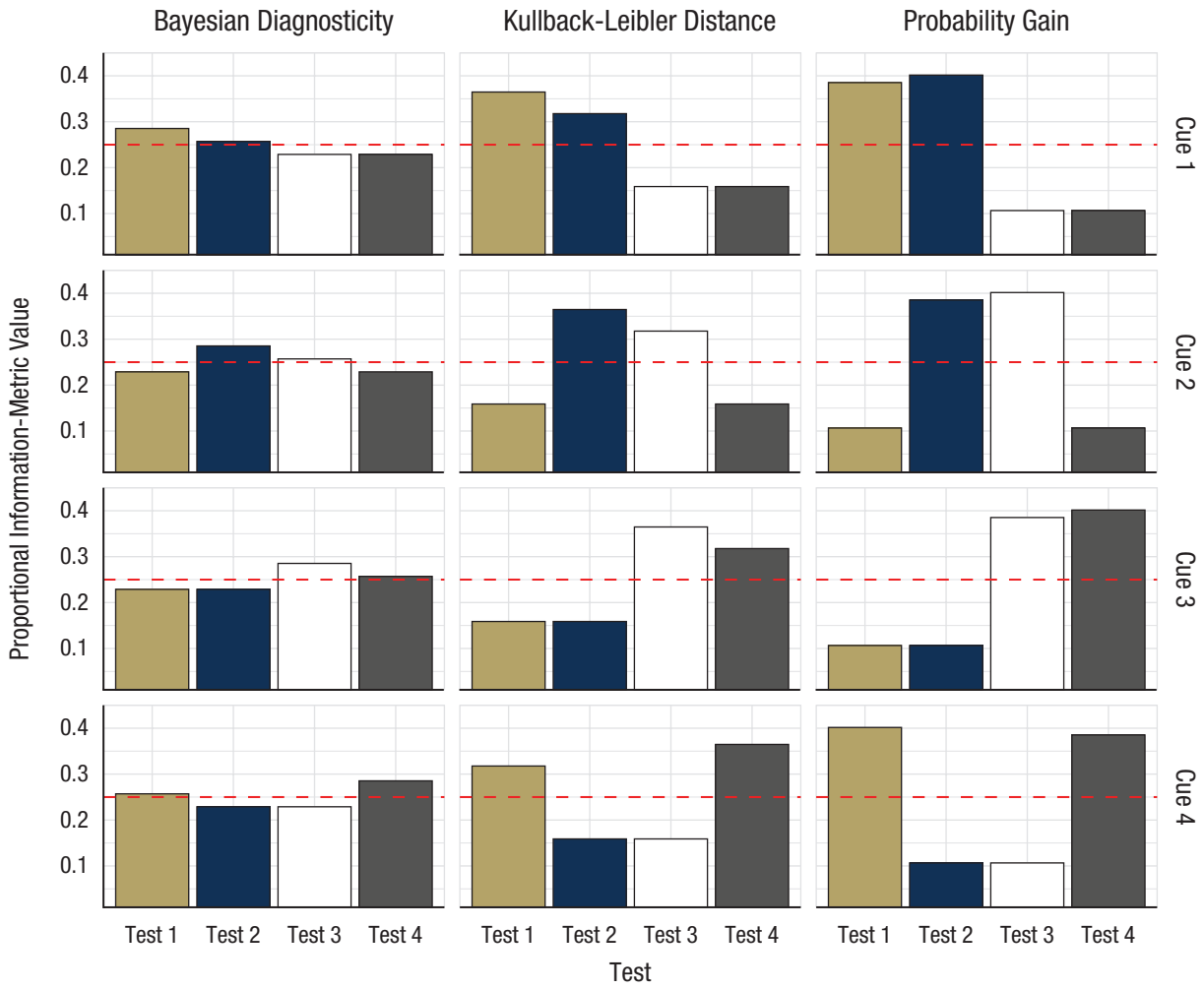

Fig. 2. Relative value of three information metrics for three tests, separately for each presenting cue. This figure illustrates the predicted preference for tests, as determined by the change in information metrics after the presenting cue is interpreted. The dashed line represents the relative value for all tests before any cue is presented.

represent the conditional probability that each of three outcomes will result from a medical test, given that the patient was experiencing one of the four disease states. A critical feature of the tests was that their informative value was constrained to be equal when a decision-maker believed that all four hypotheses were equally likely. We used a modified form of Bayesian diagnosticity (see Equation 2) to operationalize the informative value of information sources and held this value constant across all tests (diagnosticity = 6.38):

$$\text{diagnosticity}(T) = \sum_{j=1}^{3} \left(P(d_j) \times \frac{\sum_{i=1}^{4} \max\left(\frac{P(d_j|H_i)}{P(d_j|H_0)}, \frac{P(d_j|H_0)}{P(d_j|H_i)} \right)}{4} \right) \quad (2)$$

$$\text{Kullback-Leibler distance}(T) = \sum_{j=1}^{3} \left(P(d_j) \times \sum_{H_i} P(H_i|d_j) \log_2 \frac{P(H_i|d_j)}{P(H_i)} \right) \quad (3)$$

$$\text{probability gain}(T) = \left(\sum_{j=1}^{3} P(d_j) \times \max\left(P(H_i \mid d_j) \right) \right) - P(H_{\max}) \quad (4)$$

The modified form was needed given that Bayesian diagnosticity is not mathematically defined for more than two hypotheses (Good, 1950; Nelson, 2005). We took the maximum likelihood for each of the four hypotheses and computed a mean, which estimates the expected maximum likelihood of acquiring a specific

Table 2. Objective-Medical-Test Statistical Environment

Hypothesis and data	Test 1	Test 2	Test 3	Test 4
Hypothesis 1				
Outcome 1	.90	.05	.05	.05
Outcome 2	.05	.50	.50	.50
Outcome 3	.05	.45	.45	.45
Hypothesis 2				
Outcome 1	.05	.90	.05	.05
Outcome 2	.50	.05	.50	.50
Outcome 3	.45	.05	.45	.45
Hypothesis 3				
Outcome 1	.05	.05	.90	.05
Outcome 2	.50	.50	.05	.50
Outcome 3	.45	.45	.05	.45
Hypothesis 4				
Outcome 1	.05	.05	.05	.90
Outcome 2	.50	.50	.50	.05
Outcome 3	.45	.45	.45	.05
Diagnosticity	6.38	6.38	6.38	6.38
Probability gain	.21	.21	.21	.21
Kullback-Leibler distance	0.50	0.50	0.50	0.50

Note: Values represent the probability of observing an outcome from each test given the underlying disease hypothesis.

piece of data (test outcome; d_j). Thereafter, diagnosticity was computed normally by weighting the maximum likelihood by the probability of the test outcomes before summing those values for all three possible test outcomes. We also computed the value of each test using Kullback-Leibler distance (Equation 3) and probability gain (Equation 4). The value of each test as computed by each of the three metrics is shown in Table 2. All information metrics agree that the tests are equally valuable when belief in the four hypotheses is equal.

The diagnostic value of each test changed drastically with the onset of a presenting cue. Specifically, the medical tests were designed to map closely to the cue configuration presented in Table 1. Cue 1, for example, was strongly associated with Hypothesis 1, whereas Hypothesis 2 was its nearest competitor—accounting for 50% and 30% of cases presenting Cue 1, respectively. Test 1 became the most diagnostic test given Cue 1 presentation (diagnosticity = 7.28), and Test 2 was its closest competitor (diagnosticity = 6.56). Meanwhile, the diagnostic value of Tests 3 and 4 dropped from the controlled initial value (diagnosticity = 5.84). Figure 2 illustrates a staggered design in which this pattern persists across cues and tests, where Cue 2 increases the diagnostic values of Tests 2 and 3, and so on. We normed the value of each information metric so they could be plotted together with values that range from 0 to 1. The general patterns for changes in Bayesian diagnosticity, Kullback-Leibler distance, and probability gain are nearly identical. Note that the posterior diagnostic values are equivalent across all cue conditions; only the identities of the most useful tests change (see Table A1 in the Appendix).

Procedure

We used an experimental paradigm that we refer to as the *medical-diagnosis game* (Illingworth & Thomas, 2015), a gamified, forced-choice, simulated diagnosis task modeled after experience-based category-learning paradigms (e.g., Hoffman & Rehder, 2010; Posner & Keele, 1968). The medical-diagnosis game consists of two phases. The purpose of the first phase of the game is to facilitate participants' learning of the probabilistic relation between disease states, presenting cues, and test outcomes (Tables 1 and 2). This learning phase was completed over 24 blocks of 20 trials (each disease hypothesis was equally represented in the 20 trials), resulting in 480 total learning trials. Learning was designed to be active, as participants were required to guess an answer starting with the first trial despite possessing no knowledge of the ecological structure of the data environment. The gamification of the task incentivized accurate medical diagnosis; participants were instructed to use Phase 1 trials to accumulate as many points as possible. Participants were awarded 1,000 points (in the form of game dollars) for correct diagnoses. There was no reward or penalty for incorrect responses.

Presenting symptoms took the form of four common medical conditions: fever, rash, migraine, and ache. These labels were randomly assigned to the cues shown in Table 1. Similarly, medical tests consisted of common medical procedures: computed tomography scans, chest cavity X-rays, bacterial cultures, and abdominal MRIs. The stimuli representing medical test results were circular black-and-white images with distinct features that signaled a positive, neutral, or negative outcome (Outcomes 1–3 in Table 2). Test labels and their corresponding images were randomly assigned to each test appearing in Table 2 for each participant.

Each trial involved a new patient in need of a medical diagnosis whose ailment was one of four fictitious, mutually exclusive diseases: metalytis, zymosis, gwaronia, or descolada. These labels were randomly assigned to each disease hypothesis in Table 1 for each participant. At the start of a trial, only the presenting cue and the response scale were visible (see Fig. 3). After 200 ms elapsed, the outcomes of all tests were presented simultaneously. Participants submitted their diagnoses at their own

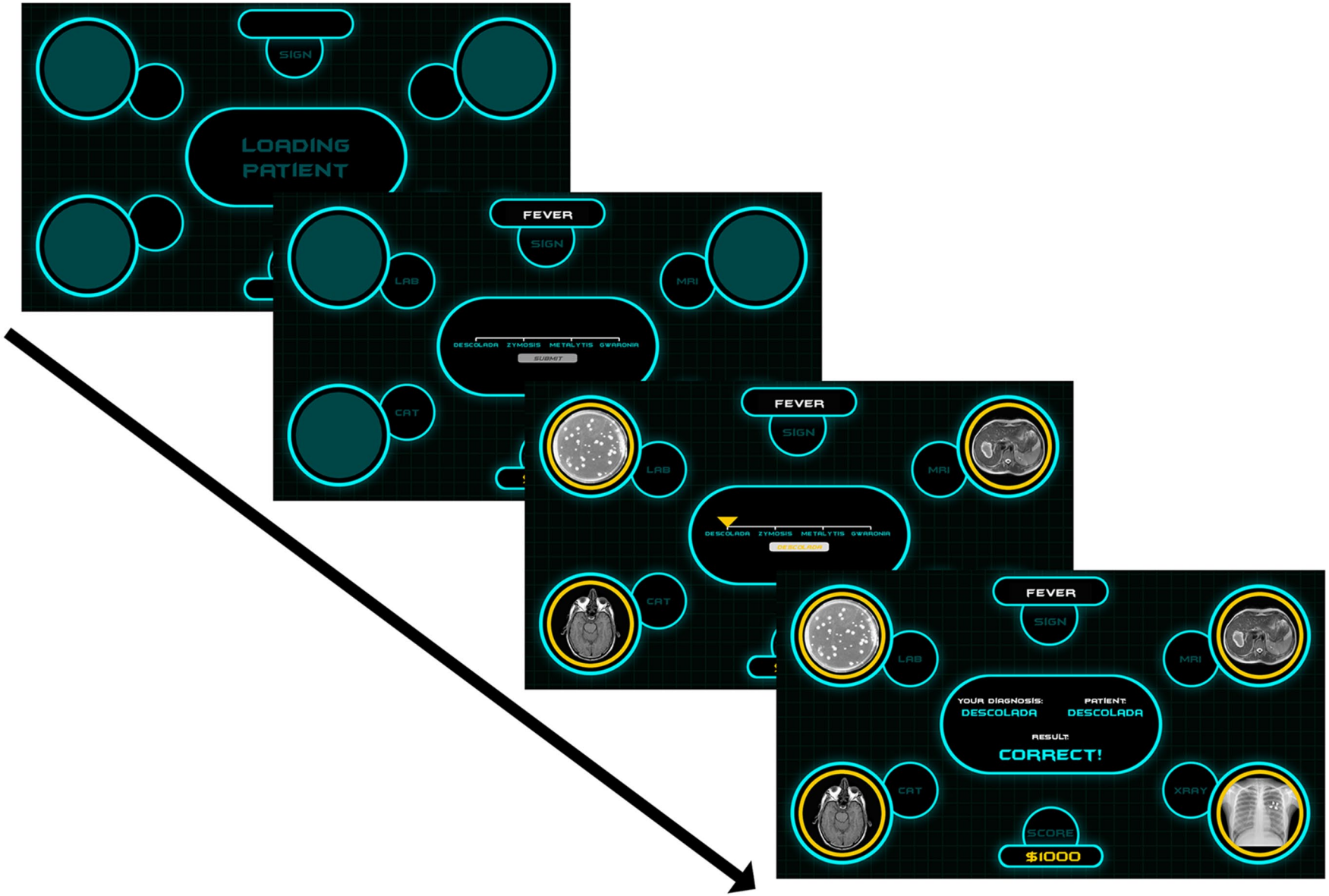

Fig. 3. Illustration of a learning trial. At the start of each trial, the presenting cue (fever, rash, migraine, or ache) appeared at the top of the screen, and the response scale (on which participants made one of four medical diagnoses: metalytis, zymosis, gwaronia, or descolada) appeared in the middle. The outcomes of four tests were presented simultaneously after 200 ms: computed tomography (CAT) scans, chest cavity X-rays, bacterial cultures (LAB), and abdominal MRIs. Feedback was provided after each diagnosis, and the updated number of points accumulated was shown at the bottom of the display. Test trials differed in that results appeared only after participants clicked on the desired test. Feedback was provided for learning trials only.

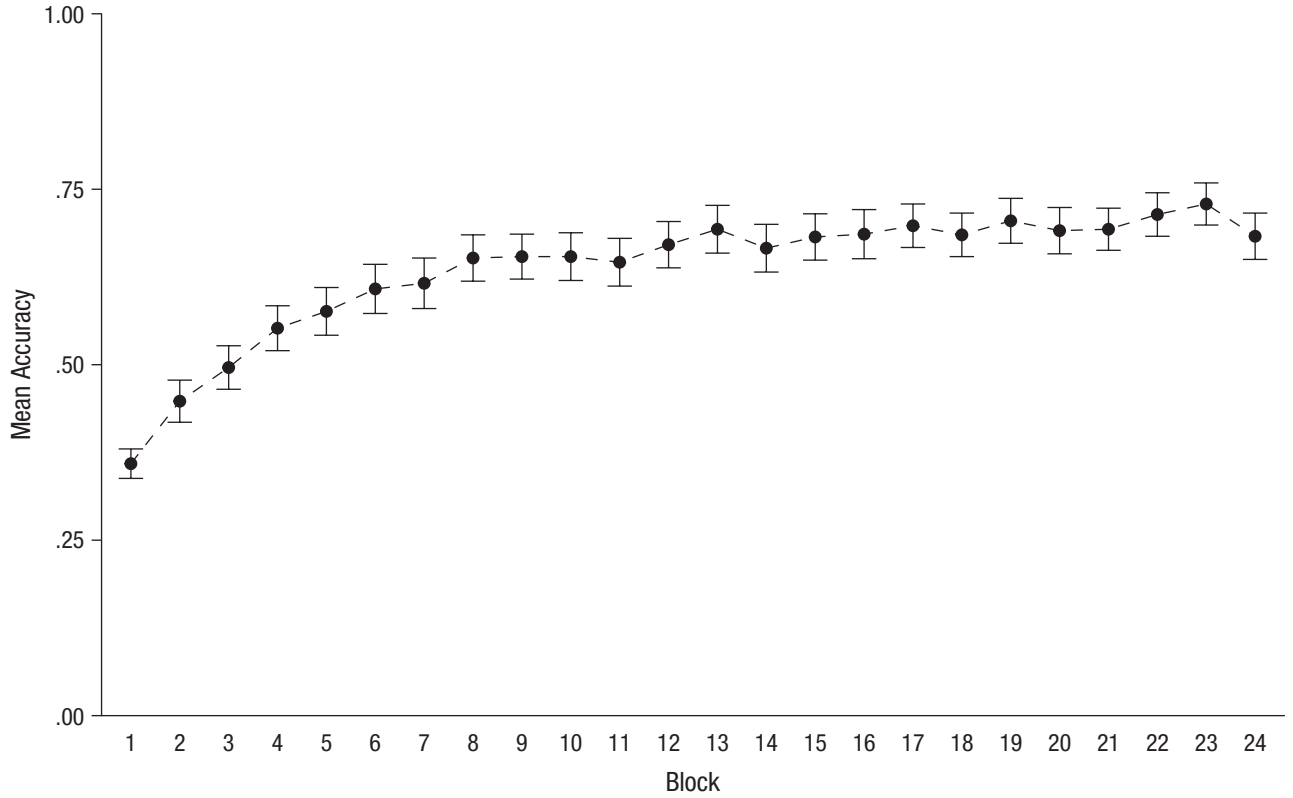

Fig. 4. Mean proportion of correct responses in each block of the learning phase. Error bars represent standard errors of the mean.

discretion. Diagnoses could be changed indefinitely until the participant submitted their response. Feedback was provided after each diagnosis; the word "CORRECT!" appeared after participants diagnosed the patient with the appropriate disease, and "INCORRECT" appeared after an erroneous diagnosis. The number of points accumulated was updated on the feedback screen and was visible at the bottom of the display. No criterion was used to denote learning. All participants moved on to the test phase of the experiment at the completion of the learning phase.

The test phase was completed over four blocks of 20 trials each, for a total of 80 test trials. The disease hypotheses were equally represented within each block of test trials. Test-phase trials differed from those of the learning phase in that test results did not automatically appear after a short latency. Rather, participants had to explicitly click on the desired test's circular widget before the corresponding outcome would appear. Tests were, thus, selected sequentially. The number of tests viewed was left to the participant's discretion, and termination of search (i.e., submission of their diagnosis) could occur after viewing between none and all of the test outcomes. Each test trial was self-paced, and as was the case in learning trials, diagnoses could be changed indefinitely until the participant submitted their response. Feedback was withheld from participants during the test phase of the experiment.

Results

Learning phase

Before testing our main hypotheses, we analyzed learning-phase behavior to assess how well participants learned to accurately diagnose patients in this task. A logistic regression evaluated how well accuracy—coded 1 for correct responses and 0 for incorrect responses—was predicted by trial block. Trial block was found to be predictive of learning phase accuracy, $\chi^2(23, N = 56) = 888.60$, $p < .001$, suggesting that performance improved over the course of the 24 blocks of learning trials. Participants were more likely to submit a correct diagnosis in the final block than they were in the first block (odds ratio [*OR*] = 5.22, $p < .001$, 95% confidence interval [CI] = [5.03, 5.41]). This result is illustrated in Figure 4.

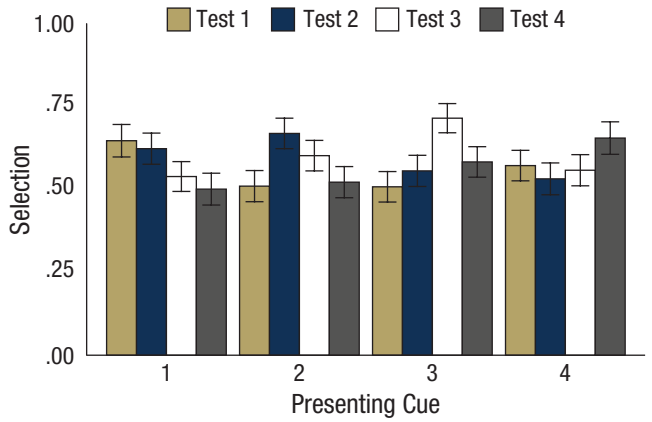

Fig. 5. Mean proportion of trials on which each test was selected, separately for each presenting cue. Error bars represent standard errors of the mean.

Test phase

We evaluated the hypothesis-guided-search principle by analyzing medical-test-selection behavior. Selection was coded 1 if a test was selected in the trial and 0 if it was not selected. A binomial logistic regression analysis evaluated whether selection could be predicted by cue, test, and the interaction of the two factors. Presenting cue, $\chi^2(3, N = 56) = 26.40$, $p < .001$, and test, $\chi^2(3, N = 56) = 106.60$, $p < .001$, were both found to be predictive of selection behavior. However, both results were qualified by a Presenting Cue × Test interaction, $\chi^2(9, N = 56) = 1{,}233.00$, $p < .001$. This omnibus interaction is illustrated in Figure 5, in which it is clear that the frequency with which each test was selected varied with respect to the information gleaned from the presenting cue.

We carried out additional analyses of testing behavior to follow up on the omnibus interaction by evaluating selection of each test as an independent outcome variable, which allowed us to determine how selection of each test changed with respect to the presenting cue. The α level was modified with a Bonferroni correction to account for this family of analyses ($\alpha = .05/4 = .0125$). Test 1 selection was predicted by presenting cue, $\chi^2(3, N = 56) = 42.63$, $p < .001$. Participants' preference for Test 1 changed in the predicted fashion; Test 1 was more likely to be selected after Cue 1 was observed rather than Cue 2 ($OR = 2.59$, $p < .001$, 95% CI = [2.37, 2.82]), Cue 3 ($OR = 2.60$, $p < .001$, 95% CI = [2.37, 2.83]), or Cue 4 ($OR = 1.74$, $p < .001$, 95% CI = [2.37, 2.82]).

The odds of selecting Test 2, $\chi^2(3, N = 56) = 64.88$, $p < .001$; Test 3, $\chi^2(3, N = 56) = 83.84$, $p < .001$; and Test 4, $\chi^2(3, N = 56) = 53.5$, $p < .001$, were predicted by the presenting cue. The influence of presenting cue for Tests 2, 3, and 4 mirrored those of Test 1. A summary of the model statistics is reported in Table 3. These four analyses link each test to the cue predicted to result in the greatest rate of selection, which is consistent with the HyGene simulation and the information metrics discussed above. Moreover, they illustrate the graded nature of test selection, which is expected if the strength of belief drives performance in information-foraging tasks. For example, Test 4 was selected at the highest rate after Cue 4 and attracted the second highest rate of selection after presentation of Cue 3, where it also experienced a boost in diagnostic value over the controlled baseline.

We created a preference score to account for the order in which tests were selected in each trial. Order was reverse scored so that tests selected first were scored 4, tests selected second were scored 3, tests selected third were scored 2, tests selected fourth were scored 1, and unselected tests remained at 0. These values were normalized so that the total preference across the four tests summed to 1 (Rehder & Hoffman, 2005). A repeated measures analysis of variance detected a main effect of test, $F(3, 55) = 21.92$, $p < .001$, $\eta^2 = .05$, which was qualified by a significant interaction between test and presenting cue, $F(9, 55) = 61.84$, $p < .001$, $\eta^2 = .41$. Figure 6 illustrates the results of this analysis.

The pattern for test preference was nearly identical to that of test selection, indicating that the predicted relation between presenting cue and test selection also manifested in the order in which tests were selected. This was tested statistically by evaluating the simple effects for each test. Presenting cue significantly affected preference for Test 1, $F(3, 55) = 66.55$, $p < .001$, $\eta^2 = .08$; Test 2, $F(3, 55) = 58.13$, $p < .001$, $\eta^2 = .04$; Test 3, $F(3,$

Table 3. Summary of Analyses of Selection Behavior for Tests 2, 3, and 4

Analysis and cue	*OR*	*p*	95% CI
Test 2			
Cue 1	1.52	< .001	[1.28, 1.75]
Cue 3	2.31	< .001	[2.09, 2.54]
Cue 4	2.77	< .001	[2.54, 3.00]
Test 3			
Cue 1	3.37	< .001	[3.14, 3.60]
Cue 2	2.24	< .001	[2.01, 2.46]
Cue 4	3.01	< .001	[2.78, 3.24]
Test 4			
Cue 1	2.99	< .001	[2.75, 3.22]
Cue 2	2.64	< .001	[2.42, 2.88]
Cue 3	1.69	< .001	[1.46, 1.92]

Note: The odds ratio (*OR*) should be interpreted as the degree to which the cue serving as the intercept (e.g., Cue 2 for Test 2) resulted in a greater likelihood of selection for the analyzed test. CI = confidence interval.

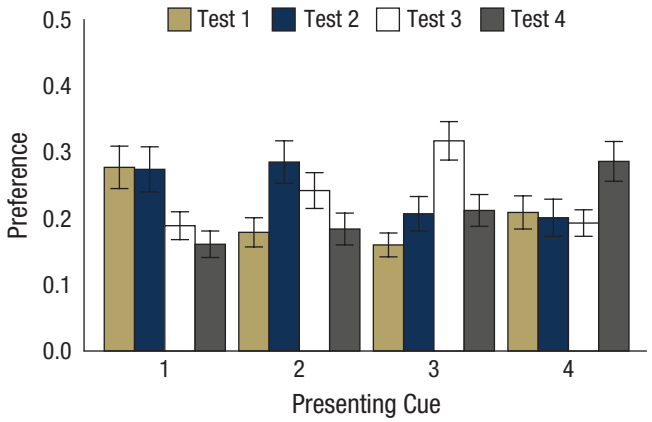

Fig. 6. Order in which each test was selected (preference score), separately for each presenting cue. Error bars represent standard errors of the mean.

55) = 86.98, $p < .001$, $\eta^2 = .11$; and Test 4, $F(3, 55) = 79.98$, $p < .001$, $\eta^2 = .07$. Post hoc tests are reported in Table 4.

We further explored the observed search behavior by examining decisions to terminate search and how they related to task performance. To do so, we first computed the objective maximum Bayesian posterior belief for every trial for each participant using the presenting cue and the test outcomes observed as recorded by their search behavior. The relations between participants' mean accuracy, mean tests selected, and mean maximum posterior were estimated by Kendall's τ. Significant and positive correlations were found between accuracy and total tests selected ($r = .30$, $p < .05$), as well as accuracy and maximum posterior belief ($r = .29$, $p < .05$).

Our observation that the relations between both total number of tests selected and accuracy and between maximum posterior belief and accuracy, but not with each other, indicates that the dynamics of the observed search behavior are likely to have been data driven. This suggests that participants were responsive to the data environment. To test this notion, we created a simple within-subjects code for trials in which the result of the first test selection was confirmatory (increased belief in the disease hypothesis that was most strongly indicated by the presenting cue) and those trials in which the result was not confirmatory. A paired-samples *t* test showed that participants selected significantly more tests when the first observed result was not confirmatory ($M = 2.47$, $SE = 0.13$) than when it was confirmatory ($M = 1.90$, $SE = 0.14$), $t(52) = 4.61$, $p < .001$, $d = 0.55$. This result suggests that participants with more uncertain beliefs, as can be inferred from an objective standard of the data observed, engaged in more search than those with more certain beliefs.

Discussion

Numerous information-search and information-valuation theories assume that beliefs in hypotheses drive information selection. The patterns of information acquisition observed in our experiment show that belief updating, not other statistical properties of the data environment, guided test selection. The experiment's key design feature was controlling the (objective) initial informative values of the available information sources, the base rates of the hypotheses, and the prevalence of the test outcomes. Our results contribute to the information-search literature by demonstrating the role of pre-search processes (memory retrieval and belief updating) in information search, thus providing a cognitive-process account for the manifestation of belief relevant to theories of information valuation and search. We note, however, that the generalizability of these findings is limited given that our participants were sampled entirely from a pool of college students in the United States.

The results are complementary to previous work that focused on the quality of information attended to by decision-makers. For example, Illingworth and Thomas (2015) found evidence of sensitivity to informative value during data acquisition in a sequential hypothesis-testing task. Their study showed that participants exhibited a strong preference for highly informative tests regardless of the shape of the cost distributions (i.e., equal costs vs. unequal costs) and the distinctiveness of the test outcomes. Nelson and colleagues (Nelson, 2005; Nelson et al., 2010) have repeatedly found evidence supporting

Table 4. Results of Tukey's Honestly-Significant-Difference Analyses for Simple Effects of Preference

Analysis and cue comparison	*q*	*z*	*p*
Test 1			
2 vs. 1	0.09	11.25	< .001
3 vs. 1	0.11	13.25	< .001
4 vs. 1	0.07	7.59	< .001
Test 2			
1 vs. 2	0.02	2.05	.17
3 vs. 2	0.08	9.18	< .001
4 vs. 2	0.08	9.83	< .001
Test 3			
1 vs. 3	0.12	14.09	< .001
2 vs. 3	0.07	8.46	< .001
4 vs. 3	0.12	13.72	< .001
Test 4			
1 vs. 4	0.12	14.65	< .001
2 vs. 4	0.10	12.87	< .001
3 vs. 4	0.07	9.07	< .001

sensitivity to informative value during data acquisition. Whereas Nelson and colleagues have focused on the fit of different metrics of information value to information-consumption patterns, the metrics investigated in their studies include a parameter representing the strength of belief in hypotheses. Our study expanded on their work by examining the processes that precede the selection of information sources.

This experiment's results were also consistent with previous findings that demonstrated the conditions under which participants engage in pseudodiagnostic or diagnostic search. Dougherty et al. (2010) reported evidence suggesting that pseudodiagnostic search was a direct by-product of retrieval dynamics and belief updating and that a preference for pseudodiagnostic tests was more likely when there was a strong belief in a single hypothesis. The presenting-cue statistical environment detailed in Table 1 elicited diagnostic search behavior by always promoting increased belief in different sets of two disease hypotheses—the updated beliefs then influenced test preference.

In most studies of information search, manipulations are conducted at the level of the information source, where the data-acquisition patterns reveal something about the information-search rule or strategy used to evaluate the sources' usefulness. Our approach was to control the initial, objective informative value of all the information sources as well as the base rates of all observable data (presenting symptoms and test outcomes) and hypotheses. Although simple, this design allowed us to infer that the changes we observed in test preference were most likely driven by the strength of belief in the hypotheses. This degree of control has illustrated, in part, the care necessary in the construction of information environments to address basic predictions of general cognitive search.

The importance of advancing theory relevant to information-source valuation consistent with the cognitive mechanisms that motivate or guide search behavior is an issue discussed previously by Coenen et al. (2018). Process models, such as HyGene, can serve as valuable tools for understanding pre-search processes and can be extended to formalize information-source valuation (e.g., Dougherty et al., 2010) and decisions to terminate search. Such process models provide heuristic value for directing future work, particularly when the objective information metrics disagree with the predictions of cognitive frameworks such as HyGene.

To better illustrate this point, we ran an additional HyGene simulation with four presenting cues and four tests with three outcomes each. In the data environment implemented in this simulation, the presenting cues did not change the tests' relative informative values across the different information metrics (Fig. 7a).

Although the information metrics use the full set of hypotheses in their calculations, HyGene has access only to those hypotheses explicitly generated via recall to "calculate" its information preferences. Given only the stochastic retrieval dynamics, different simulated participants base their information preferences on different sets of hypotheses that vary in both number and composition. Figure 7b illustrates the model's test preferences, broken down by the number of hypotheses generated by the simulated participants. One can see that the pattern of test preference differs as a function of the number of hypotheses considered by the simulated participants. The patterns of test preference also deviate from the (objective) information metrics. Several pre-search processes within the HyGene model influence the number of hypotheses generated that affect test preferences, including individual differences (retrieval-termination rules and working memory capacity) and task characteristics (external time pressure and dual task). For example, participants with low working memory capacity, under dual-task conditions or under time pressure, would exhibit patterns of test preference similar to when the number of hypotheses considered is low (Fig. 7b, top two rows). In contrast, participants with high working capacity, undivided attention, and less time pressure would exhibit patterns of test preference similar to when the number of hypotheses considered is high (Fig. 7b, bottom two rows). It should be noted that HyGene's predictions are likely not unique in that many cognitive-process models could make similar predictions if their retrieval processes were augmented to constrain information search (for an instantiation of HyGene in the ACT-R architecture, see Dimov, 2018).

Hypothesis-guided search is a ubiquitous behavior exhibited by people across a myriad of settings. Poletiek (2001) lists numerous behaviors (e.g., orienting a glance to evaluate expectations of surrounding objects, uttering sounds while learning a language to assess one's mastery of novel phonemes, solving problems in unique ways to observe their impact on the world) that are, in essence, instances in which expectations guide how people extract information from their surroundings. These behaviors are manifested in highly complex environments in which numerous ecological or individual factors may impact the incentives, risks, and costs inherent to data acquisition (cf. Meier & Blair, 2013). Our goal has been to carefully construct the simplest of scenarios to evaluate the predicted relation between belief and information search. Future research is necessary to further explore whether and how this process can account for the numerous outstanding questions that remain in the domain of human inquiry (Coenen et al., 2018).

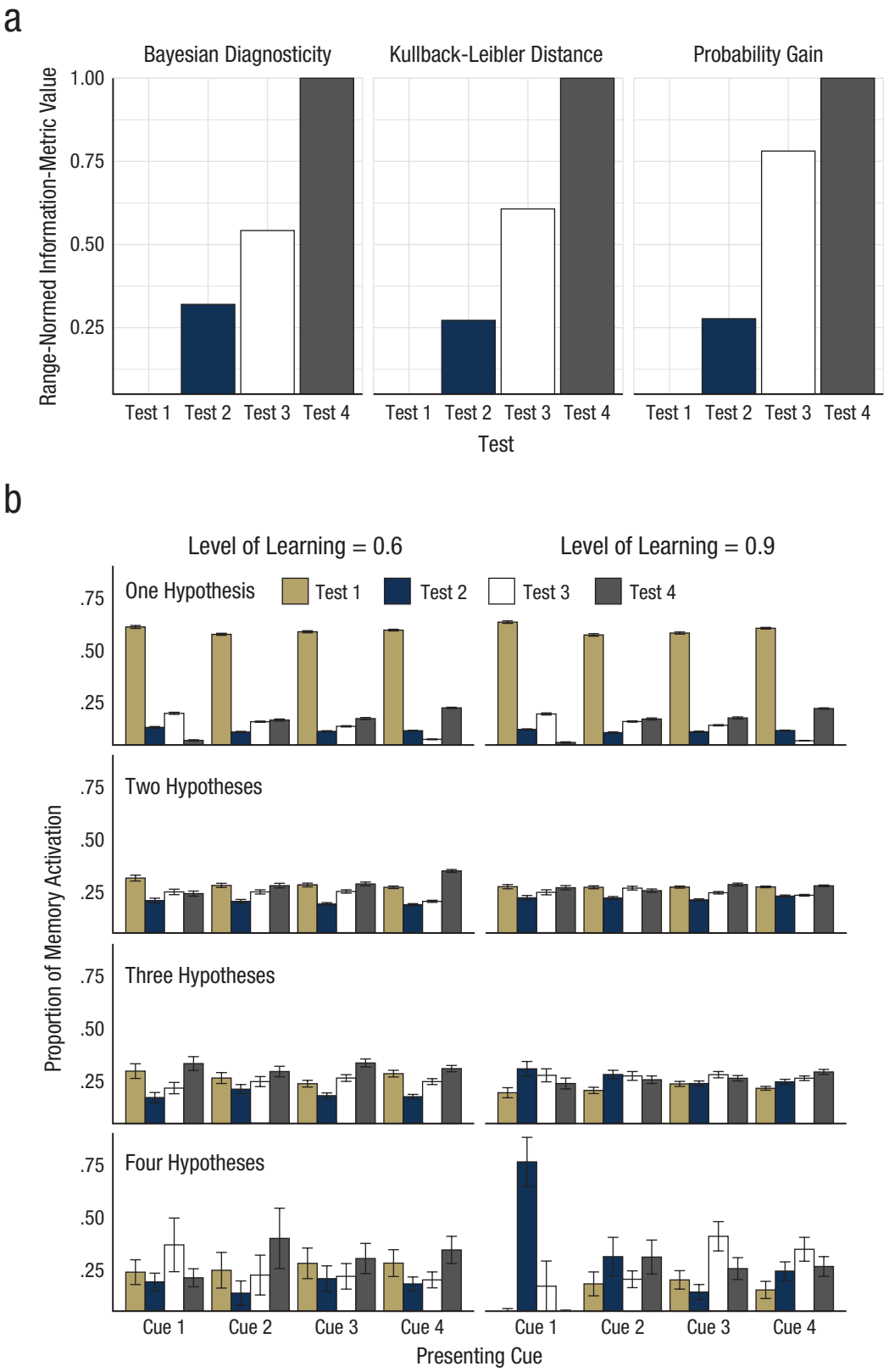

Fig. 7. HyGene behavior in an alternative data environment with four presenting cues and four tests with three outcomes each. Range-normed values of three objective information metrics (a) are shown for each test. The model-estimated proportion of memories activated by each presenting cue (b) is shown for each level of learning and number of hypotheses generated by the model. Error bars represent standard errors of the mean.

Appendix

Table A1. Test Value of Three Objective Information Metrics After Observation of Each Presenting Symptom

Presenting cue and test	Diagnosticity	Kullback-Leibler distance	Probability gain
Cue 1			
Test 1	7.28	0.62	.24
Test 2	6.56	0.55	.25
Test 3	5.84	0.27	.07
Test 4	5.84	0.27	.07
Cue 2			
Test 1	5.84	0.27	.07
Test 2	7.28	0.62	.24
Test 3	6.56	0.55	.25
Test 4	5.84	0.27	.07
Cue 3			
Test 1	5.84	0.27	.07
Test 2	5.84	0.27	.07
Test 3	7.28	0.62	.24
Test 4	6.56	0.55	.25
Cue 4			
Test 1	6.56	0.55	.25
Test 2	5.84	0.27	.07
Test 3	5.84	0.27	.07
Test 4	7.28	0.62	.24

Transparency

Action Editor: Sachiko Kinoshita
Editor: Patricia J. Bauer
Author Contributions

Both authors contributed to the study concept, study design, and simulation. D. A. Illingworth conducted testing and data collection, analyzed and interpreted the data, and drafted the manuscript. R. P. Thomas provided critical revisions to the manuscript. Both authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

All data, analysis codes, and codes to generate the materials have been made publicly available via OSF and can be accessed at https://osf.io/432u7/. The design and analysis plan for the study were preregistered at https://osf.io/67z5v/. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

David A. Illingworth https://orcid.org/0000-0003-4017-5801

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976211043425

References

Buttaccio, D. R., Lange, N. D., Thomas, R. P., & Dougherty, M. R. (2015). Using a model of hypothesis generation to predict eye movements in a visual search task. *Memory & Cognition*, *43*(2), 247–265.

Buttaccio, D. R., Lange, N. D., Thomas, R. P., & Dougherty, M. R. (2018). Does constraining memory maintenance reduce visual search efficiency? *Quarterly Journal of Experimental Psychology*, *71*(3), 605–621.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2018). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin and Review*, *26*, 1548–1487.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942.

Dimov, C. M. (2018). How to implement HyGene into ACT-R. *Journal of Cognitive Psychology*, *30*(2), 163–176.

Dougherty, M., Thomas, R., & Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 52, pp. 299–342). Academic Press.

Evans, J. S. B., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103*(2), 356–363.

Good, I. J. (1950). *Probability and the weighting of evidence*. Charles Griffin.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science*, *30*(1), 3–41.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319–340.

Illingworth, D. A., & Thomas, R. P. (2015). Price as information: Incidental search costs affect decisions to terminate information search and valuations of information sources. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 225–229.

Johnson-Laird, P. N., & Byrne, R. M. (1991). *Deduction*. Erlbaum.

Kirby, K. N. (1994). False alarm: A reply to Over and Evans. *Cognition*, *52*(3), 245–250.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228.

Manktelow, K. I., & Over, D. E. (1990). Deontic thought and the selection task. *Lines of Thinking*, *1*, 153–164.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191–215.

Meier, K. M., & Blair, M. R. (2013). Waiting and weighting: Information sampling is a balance between efficiency and error-reduction. *Cognition*, *126*(2), 319–325.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact and information gain. *Psychological Review*, *112*, 979–999.

Nelson, J. D., McKenzie, C. R. M., Cotrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*(7), 960–969.

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review*, *103*(2), 381–391.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*, 643–675.

Poletiek, F. H. (1995). Testing in a rule discovery task: Strategies of test choice and test result interpretation. In J.-P. Caverni, M. Bar-Hillel, H. Barron, & H. Jungermann (Eds.), *Contributions to Decision Research-I* (pp. 335–350). Elsevier.

Poletiek, F. H. (2001). *Hypothesis-testing behaviour*. Psychology Press.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*(1), 155–185.