Adaptive information source selection during hypothesis testing

Andrew T. Hendrickson (drew.hendrickson@adelaide.edu.au)
Amy F. Perfors (amy.perfors@adelaide.edu.au)
Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, Level 4 Hughes Building,
University of Adelaide, SA 5005, Australia

Abstract

We consider how the information sources people use to test hypotheses change as the sparsity of the hypotheses – the proportion of items in the hypothesis space they include – changes. Specifically, we focus on understanding how requests for positive and negative evidence, which have been shown to be sensitive to hypothesis sparsity (Hendrickson, Navarro, & Perfors, in prep), are influenced by requests for specific instances, which show a positive bias and less sensitivity to sparsity (Markant & Gureckis, 2013). We find that people modify their information requests as a function of the sparsity of the hypotheses and they do so in this task primarily by by manipulating the rate of requesting positive and negative evidence. Furthermore, by simulating the set of possible remaining hypotheses, we find that people were most likely to select the information source that maximized the expected reduction in uncertainty across hypotheses. We conclude by discussing the implications of these results for models of hypothesis testing.

Keywords: Confirmation bias; positive test strategy; hypothesis testing; information search;

Introduction

How does a child learn what will float in water and what will sink? Certainly not by studying the calculus that underlies the equations for density, volume, and buoyancy. Instead, she might start by asking for examples of things that float and things that sink. She might also ask if specific objects float, like a small rock or a duck. If the goal of each information request is to minimize her uncertainty across her hypotheses about what objects float, then on average some types of information requests are going to be more informative than others, and which kinds of requests are better might change over the course of learning. In this current work we investigate if people are sensitive to differences in the informativeness of these types of information requests in a game-like task.

Foundational studies in the area of hypothesis testing showed that people have a propensity to fixate on one possible hypothesis and be biased to select test queries that generate positive results with respect to that hypothesis in tasks like discovering a number rule in the 2-4-6 number game (Wason, 1960; Klayman & Ha, 1989) or testing a verbal rule in the Wason card selection task (Wason, 1968; Klayman & Ha, 1987). This behavior was originally presented as a deviation from rational testing strategies since these information requests did not maximally reduce the uncertainty about which hypothesis was true given the full space of possible hypotheses (Nickerson, 1998).

The strategy of selecting tests that are likely to generate positive results is rational in certain domains (Oaksford & Chater, 1994). Positive tests are adaptive when hypotheses are *sparse* with regard to acceptance – that is, hypotheses are likely to respond YES to only a few possible queries. For instance, the category DOGS is sparse because most items are not dogs; most categories are similarly sparse. In situations with sparse hypotheses, generating positive tests with regard to a small subset of hypotheses (or a single, current one) will eliminate a high proportion of potential hypotheses and therefore minimize uncertainty about the true hypothesis (Navarro & Perfors, 2011).

Are people sensitive to hypothesis sparsity? Some evidence (Figure 1), from studies in which information requests were limited to *evidence requests* – random examples of either positive or negative evidence – has shown the proportion of positive evidence requests increases as sparsity increases (Langsford, Hendrickson, Perfors, & Navarro, in press; Hendrickson et al., in prep). In the converse case, when people are only permitted to ask about specific examples (*instance requests*), people are also sensitive to sparsity (Markant & Gureckis, 2013). In both cases, there was an additional bias toward preferring positive tests, although the size of the bias was much stronger for instance requests.

None of these studies, however, matches the task of a child learning what floats in water. Allowing only evidence requests corresponds to preventing the child from asking about a specific item and allowing only instance requests prevents asking for an example of something that floats or sinks. More generally, they do not map onto the real-world tasks learners face where they can make both types of information requests and must choose between them. In the current study we present a task where people were asked to find one correct hypothesis with the option of making an instance request or an evidence request of their choosing. We manipulated the sparsity of the hypotheses and also analyzed the utility of the information requests to evaluate whether people were making choices sensibly. We find that people are sensitive to hypothesis sparsity, particularly when it comes to making evidence requests, and in general are sensitive to the utility of the information in choosing between evidence and instance requests.

Experiment

This experiment evaluates what types of information participants use when attempting to search through a large set of possible hypotheses. The experiment took the form of a game loosely based on the game "Battleships", where people try to identify the location of ships, but were allowed to make information requests by selecting either a specific instance or a type of evidence. Every possible configuration of the ships was a hypothesis in the set of possible hypotheses and the

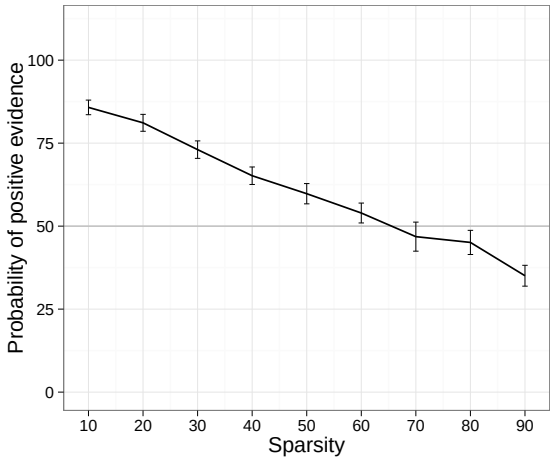

Figure 1: Data from Hendrickson et al. (in prep). People are sensitive to sparsity of the hypotheses, choosing positive evidence more often when hypotheses are more sparse in a task in which the information options were limited to postive and negative evidence requests. There is also a slight overall positive test bias, particularly when hypotheses were less sparse. Error bars indicate standard error.

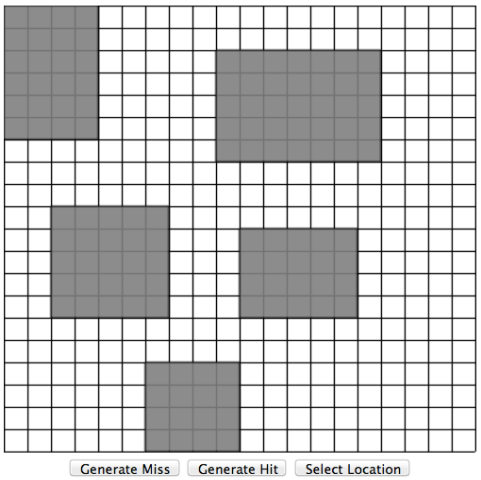

Figure 2: Sample view at the beginning of a trial from the condition in which the sparsity of each hypothesis is 30% (that is, the area covered by the five rectangles constitutes 30% of the grid). Participants could click and drag rectangles to move them within the grid or click the buttons at the bottom to generate information requests: a HIT was a request for positive evidence, a MISS was a request for negative evidence, and asking about a LOCATION was an instance request.

sparsity of those hypotheses was manipulated by changing the size of the ships.

Method

Participants We recruited 600 participants through the Amazon Mechanical Turk website, each of whom played the game three times. 47% were female and they ranged in age from 18 to 69 with a mean of 34.1. They were from 25 countries: 67% from the USA, 29% from India, with all other countries less than 1%. We report results from the 501 participants who completed all three games, made at least one information request on each game, and made no more than 100 such requests on each game. Participants were randomly assigned to one of 9 sparsity conditions, described below, with the sample size for conditions ranging from 49 to 73. The mean completion time was 16 minutes, and participants were paid US$0.60 for their time.

Materials and procedure Participants were shown a 20-by-20 grid filled with 5 rectangles ("ships") as illustrated in Figure 2. The sizes of the rectangles varied across conditions and no rectangles within a condition were exactly the same shape and size. The instructions explained that each rectangle had a hidden true position within the grid and their goal was to guess the hidden locations as closely as possible, which they could indicate by moving the visible rectangles to match those locations.

People were allowed to request information in one of three ways. They could click a button to request a piece of positive evidence (a HIT). If this option was selected, a randomly chosen grid cell that was within the true hidden location of a rectangle was marked on screen with a red box as in Figure 3. They could also request negative evidence (a MISS) in which case a random grid cell that was not occupied by the hidden location of a rectangle was shown in blue. We refer to these two types of requests as *evidence requests*. Finally, they could select a specific cell by clicking on it, and it would be labelled as either a hit or a miss depending on whether that cell was occupied by one of the hidden locations of the rectangles. We refer to this kind of request as an *instance request*.

After each information request, participants could move the rectangles within the grid. They were not permitted to make additional information requests until their current guess of the rectangle locations was consistent with all revealed information. People were asked to indicate they were done when they felt confident in their guesses about the hidden position of the rectangles. Afterwards they were shown the position of the hidden rectangles and their final guesses. They were also given a score based on the Euclidean distance between their final guesses and the actual positions, divided by the number of information requests.

There were nine experimental conditions, defined by the proportion of the grid covered by rectangles, which ranged from 10% to 90% in step sizes of 10%. Because each possible hypothesis corresponds to a configuration of rectangles, every hypothesis within a condition had the same sparsity: in the 20% condition, every valid hypothesis implies that 20% of the cells were hits, and 80% were misses. By manipulating the size of the rectangles across conditions, we were able to control the overall sparsity of the hypothesis space. Participants played three games each, all within the same sparsity condition. The first game was a practice task in which they

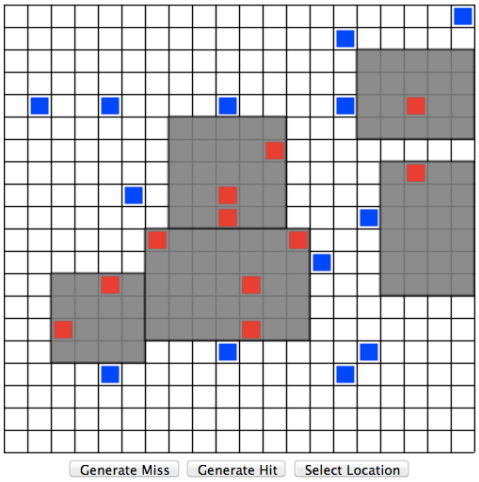

Figure 3: Sample view from the same condition as Figure 2 after 25 information requests. Requests that produced an instance within a hidden rectangle are colored red, while instances outside the hidden rectangles are colored blue. After each information request the participant had to move the rectangles into a position that is consistent with all of the information. This is one possible arrangement of rectangles consistent with all information requests.

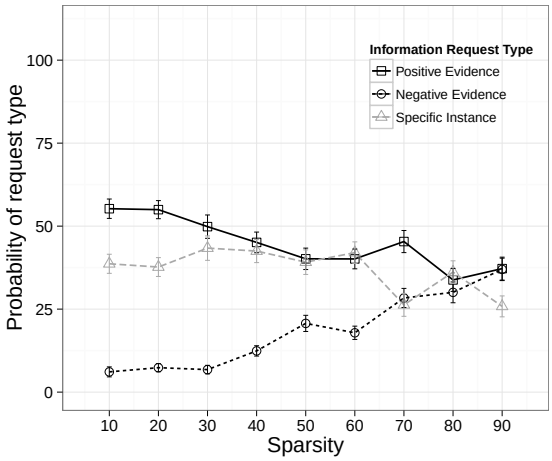

Figure 4: The proportion of information requests across hypothesis sparsity. Across all levels of sparsity participants made all three types of requests. The proportion of positive and negative evidence requests were strongly influenced by hypothesis sparsity, The effect of sparsity on on instance requests was much less strong. The solid line shows positive evidence requests, the dotted line shows negative evidence, and the dashed line shows instance requests. Error bars indicate standard error.

were instructed to try each information source at least once and familiarize themselves with the game. We analyze results from the second and third games only.

Results

Evidence requests Does the proportion of positive and negative evidence requests change as a function of the sparsity of the hypotheses? Figure 4 shows a significant effect of hypothesis sparsity on the proportion of requests for both positive evidence ($F(1,499) = 21.95, p < 0.0005$) and negative evidence ($F(1,499) = 104.6, p < 0.0005$). As hypotheses became less sparse, positive evidence requests were increasingly less likely: the slope of the line of best fit was -0.25. The pattern was the opposite for negative evidence, with more negative evidence requests occurring when the hypotheses were less sparse (the slope of the line of best fit was 0.40).

With the introduction of the possibility of instance requests, one might ask whether the relative proportion of positive to negative evidence requests parallels that found in Hendrickson et al. (in prep) and shown in Figure 1. Indeed, looking only among actions in which participants made evidence requests, shown in Figure 5, reveals a very similar pattern to what was found previously. The relationship between the proportion of positive to negative evidence requests is characterized by similar slopes of the line of best fit (0.57 in this study, 0.6 in the previous one) and intercept at 50% sparsity (0.71 in this study, 0.61 in the previous one).

Instance requests The instance requests, shown by the dashed grey line in Figure 4, were made frequently in all sparsity conditions. Though people were very sensitive to sparsity when making evidence requests, they appeared to be less so when making instance requests. Although the effect of sparsity was significant ($F(1,499) = 6.4, p = 0.01$), the slope of the line of best fit was less steep at -0.1. People were slightly more likely to ask about specific instances when the ships were smaller.

Were people selecting good instances when they requested a specific instance? As discussed by Navarro and Perfors (2011), an instance request maximizes expected information gain if the proportion of hypotheses consistent with it is 50%. If the sparsity is 20%, a random instance will be within a rectangle in an average 20% of hypotheses: selecting an instance that is within a rectangle in closer to 50% of hypotheses improves information gain. Conversely, if the sparsity is 80%, to increase information gain the learner should select an instance with a lower than average probability of a being within a rectangle. To compute the proportion of hypotheses consistent an instance it is necessary to know what hypotheses have still not yet been ruled out; this is computationally intensive because this must be calculated for every request made by every participant on every game. At each of these points we therefore simulated a pseudo-random subset of all possible valid rectangle-position hypotheses. These were used to estimate, for each information request by each participant, the probability that it was located within a rectangle across all possible rectangle-position hypotheses.[1]

[1] 20 games were simulated in most conditions due to computational complexity. Sample hypotheses were generated pseudo-

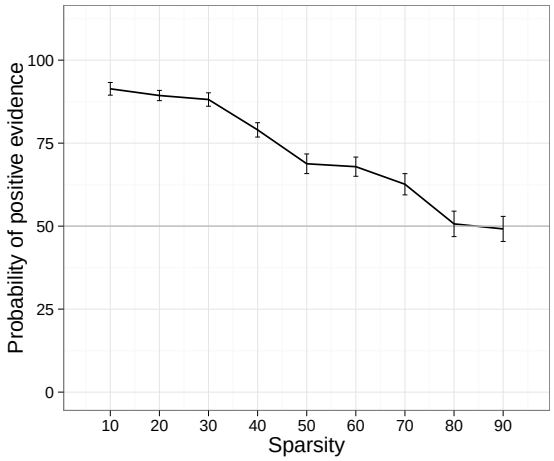

Figure 5: People are sensitive to the sparsity of hypotheses when considering only evidence requests. Even if participants had the option of requesting instances, the ratio of positive to negative evidence requests matches closely the proportion of positive requests when instance requests are not available, as shown in Figure 1 from Hendrickson et al. (in prep). Error bars indicate standard error.

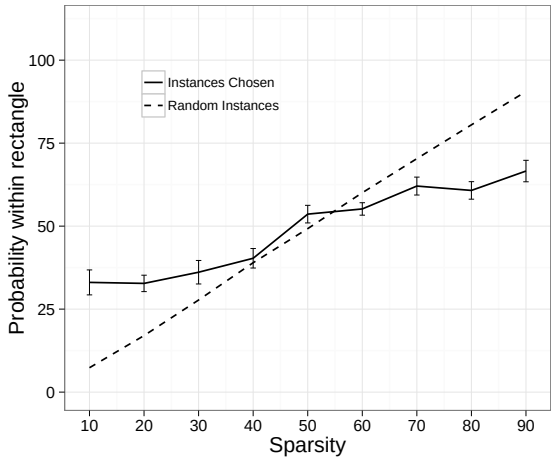

Figure 6: People do not select instances randomly. The *y* axis shows the probability than an instance selected by a participant is actually in a hidden rectangle. The dotted line reflects what would happen if people were choosing them randomly: it would parallel hypothesis sparsity. The probability of selecting a positive instance does depend on hypothesis sparsity (slope=0.46). Error bars indicate standard error.

The probability of a chosen instance being within a rectangle was strongly influenced by the sparsity of the hypothesis space ($F(1,378) = 107, p < 0.0005$, slope = 0.46) as shown by the solid line in Figure 6. Despite being influenced by the sparsity of the hypotheses, instances were not selected randomly – if they were, one would expect that the probability of any instance being located within the rectangle would be the same as the sparsity of the condition (this is reflected in the dashed line in Figure 6). We find that the instances people requested were shifted towards being equally likely to be within or outside a rectangle.

Expected utility of all information requests Why does the rate of positive and negative evidence requests depend on the sparsity of hypotheses, even when instance requests are taken into account? One possible explanation is that regardless of request type, people use a heuristic that picks information sources which are more likely to produce positive evidence when hypotheses are sparse. This heuristic predicts the linear changes seen in Figure 4 and Figure 5. An alternative explanation, first proposed by Oaksford and Chater (1994), is that people choose information sources based on how useful they expect those sources to be, and that this utility is sensitive to sparsity. One method to assess utility in this task is to look at the expected change in uncertainty about which hypothesis is correct for each information request option (for further discussion about different metrics of utility during hypothesis testing see Markant and Gureckis (2012)).

Does the relative utility of evidence requests and instance requests drive information requests beyond hypothesis sparsity? We address this by using the results of the previous simulations. For each possible information request on each action we estimated the expected reduction in the set of remaining hypotheses.[2] We then compared the utilities of each kind of information request. If people were sensitive to relative utility, one would expect that they would be more likely to choose an evidence request when it had higher utility than an instance request, and to choose an instance request when it had higher utility than an evidence request. As Figure 7 shows, that is exactly what people do. The left panel shows actions in which the utility of evidence is higher than the utility of any instance; at all sparsity levels people are more likely to make evidence requests.[3] The right panel shows the opposite situation; people were more likely to make instance requests when an instance had higher utility than either evidence type. Across all levels of hypothesis sparsity, the most frequently chosen information response option was also the highest utility option. This suggests people were sensitive to the relative utility of each information request type and used this information appropriately.

For most levels of sparsity[3] Figure 7 shows that at some times during a game instance requests had higher utility and at some times evidence requests did. How does utility depend on time, and why? Intuitively, positive evidence requests have higher utility than any instance when getting one

randomly by considering the set of possible hypotheses generated by permuting the locations of all rectangles to create a set of 120 base hypotheses then shifting each rectangle in each of those base hypotheses up to two grid cells in each direction to generate candidate hypotheses.

[2]Following Austerweil and Griffiths (2008) the expected reduction is weighted by the probability of obtaining each evidence type.

[3]For sparsity levels 40 and 50, evidence requests were never more useful than instance requests.

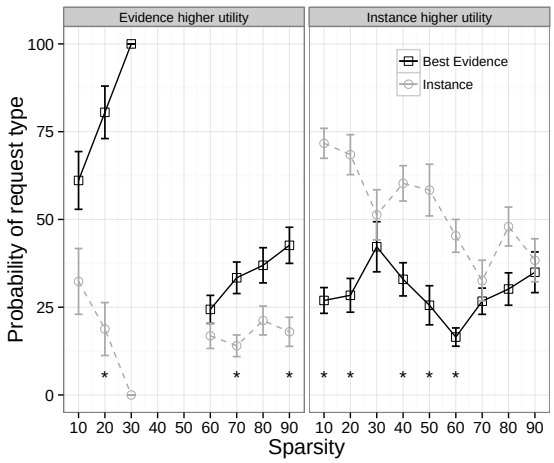

Figure 7: Proportion of information request types across sparsity. The plot is split into two panels based on which type of information request had higher utility on that action. Across all levels of sparsity people were most likely to pick the most useful information type on that action.[3] Error bars indicate standard error and stars show significant differences as indicated by a Bonferroni-corrected post-hoc test.

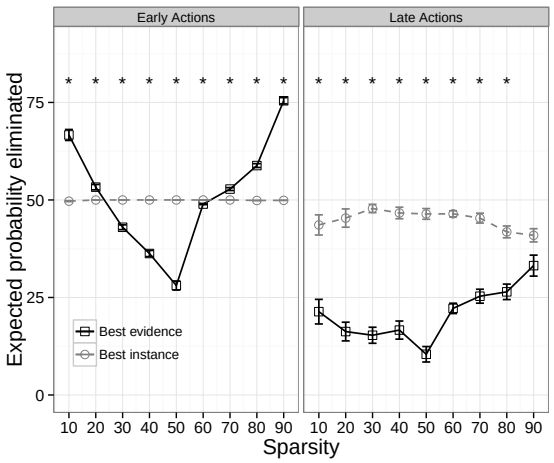

Figure 8: Utility of the best evidence request type and best instance request as a function of when in the game the request occurred. Left panel: During the first 20% of actions evidence requests had a significantly higher utility than the best instance request for extreme sparsity levels. Right: In the last 20% of actions instance requests had a higher utility than evidence requests. The black line indicates the best evidence choice, the grey line indicates the best instance choice. Error bars indicate standard error and stars show significant differences as indicated by a Bonferroni-corrected post-hoc test.

piece of positive evidence eliminates a high proportion of hypotheses. This occurs when hypotheses are sparse and most hypotheses have not been eliminated yet. Negative evidence seems to be most useful among early requests where hypotheses are not sparse at all. The left panel of Figure 8 confirms this intuition and shows that when sparsity was more extreme (either very high or very low), for early actions during a game the evidence requests had higher utility than the best instance option. However, this changed in the final actions: the best instance request had higher utility than an evidence request at all sparsity levels. Intuitively, this is because by this point individuals have narrowed down the ship locations to the point that they are deciding between just a few highly overlapping hypotheses. In this case, asking about the particular points on which those hypotheses differ is more informative than asking for evidence which may not differentiate between the remaining hypotheses.

People are sensitive to the shift in utility between request types across the course of a game. As Figure 9 shows, instance requests were more frequently made at the end of the game than in the beginning ($F(1,998) = 72, p < 0.0005$)). This effect was modulated by an interaction with hypothesis sparsity ($F(1,998) = 6.3, p = 0.01$), which suggests the difference was strongest when hypotheses were more sparse.

Discussion

In this paper we focus on understanding how sensitive people are to changes in the underlying hypothesis space when selecting different kinds of information during hypothesis testing. We find that even when they had the option of testing specific instances, they still often chose to make evidence requests asking for a HIT or a MISS. Moreover, the proportion of time that they preferred positive evidence was sensitive to sparsity. The sensitivity of evidence request type to hypothesis sparsity is one example of an underlying sensitivity to the utility of the different kinds of information requests available. People were most likely to select the information request type that was most informative about which hypothesis is correct. Moreover, this was not driven entirely by the initial hypothesis sparsity: which option had the highest utility changed throughout each game, and people were sensitive to these changes.

That said, as in Hendrickson et al. (in prep), our findings suggest that people are not perfect: we did observe a positive bias in evidence requests. As shown in Figure 5, although the proportion of positive requests is sensitive to sparsity, people have a tendency to make more positive requests than they should, especially at sparsity levels over 50%. This is not true of instance requests in all conditions, Figure 6 shows that people show a positive bias when hypotheses are sparse but a negative bias when they are not. Regardless of sparsity, the optimal instance to select is within a rectangle on 50% of hypotheses; people select instances that are less likely to be within a rectangle when hypotheses are sparse but select instances more likely to be within a rectangle when hypotheses are not sparse. One possibility for why bias among instance and evidence requests is different is that people are following different strategies for evidence and instance requests. When requesting instances, people may be aiming for the maximally-informative 50% probability instance selection by

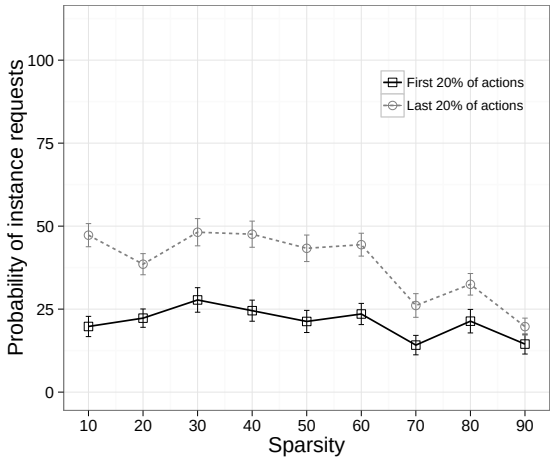

Figure 9: Proportion of instance requests across sparsity of hypotheses. Across all levels of sparsity participants made more instance requests in the last 20% requests of a game than in the first 20% requests. The black line indicates the average number of instance requests in the first 20% requests, the grey line is for the last 20% requests. Error bars indicate standard error.

only choosing such requests when they are discriminating between a few specific hypotheses, and by choosing points that would eliminate some of those hypotheses. Such a strategy, implemented non-optimally, would lead to a pattern of results as in Figure 6: the natural direction to be pulled away from optimal would be towards the informativeness of a random instance in that condition. We aim to investigate this issue further.

These results extend the growing body of work demonstrating that people are sensitive to the sparsity of the hypotheses (Langsford et al., in press; Markant & Gureckis, 2013; Hendrickson et al., in prep) in two critical ways. First, this study is the first comparison between request types within the same task and shows that sensitivity to hypothesis sparsity is not a function of having a limited set of request types. Instead, when given both evidence and instance request types, people are most likely to select the most useful type of request. Second, we believe the complexity of this task, with a large set of potential hypotheses and the full range of information request types, is the closest approximation of real-world hypothesis testing that has so far been considered in the literature of hypothesis testing. The request types and hypotheses in this "Battleships"-style game are not fundamentally different in complexity from a child learning what objects floats, a tourist learning which Australian spiders have deadly poison, or any of the other hypotheses people test and learn everyday.

If most hypotheses we encounter in the real world are sparse (Navarro & Perfors, 2011), why do people in this task bother with adjusting their responses as a function of sparsity and do not follow the simple always-select-positive-evidence heuristic? We believe the choice between many different types of information requests might drive the importance of hypothesis sparsity both in this task and in the real world. People strive to select the information request type that is most useful in reducing uncertainty (Oaksford & Chater, 1994) and the utility of request types depends heavily on the sparsity of the hypotheses (Navarro & Perfors, 2011). Therefore any computational model of hypothesis testing must move beyond simple bias-based heuristics to incorporate mechanisms that mirror human behavior and are sensitive to information utility and hypothesis sparsity when selecting among different information request options to test hypotheses.

Acknowledgments

This research was supported by ARC grant DP0773794. DJN received salary support from ARC grant FT110100431, and AP from ARC grant DE120102378.

References

Austerweil, J., & Griffiths, T. (2008). A rational analysis of confirmation with deterministic hypotheses. In *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 1041–1046). Austin, TX: Cognitive Science Society.

Hendrickson, A. T., Navarro, D. J., & Perfors, A. F. (in prep). People are sensitive to hypothesis sparsity when making information requests during hypothesis testing.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–224.

Klayman, J., & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 596-604.

Langsford, S., Hendrickson, A. T., Perfors, A. F., & Navarro, D. J. (in press). People are sensitive to hypothesis sparsity during category discrimination. In *Proceedings of the 36th annual conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Markant, D. B., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the 34th annual conference of the Cognitive Science Society* (pp. 719–724). Austin, TX: Cognitive Science Society.

Markant, D. B., & Gureckis, T. M. (2013). Changes in information search strategy under "dense" hypothesis spaces [member abstract]. In *Proceedings of the 35th annual conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120–134.

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H. (2013). assertthat: Easy pre and post assertions. [Computer software manual]. (R package version 0.1.0.99)