People are sensitive to hypothesis sparsity during category discrimination

Steven Langsford (steven.langsford@adelaide.edu.au)
Andrew T. Hendrickson (drew.hendrickson@adelaide.edu.au)
Amy Perfors (amy.perfors@adelaide.edu.au)
Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide

Abstract

Previous work has shown that the information value of requests can be manipulated by controlling the *sparsity* of hypotheses, the degree to which category members are rare or common in the domain under consideration when making those requests. However, the degree to which people are sensitive to expected information value is unknown. This study examined a binary sorting task where sparsity differed across conditions. In contrast to previous work using hypotheses representable as visual areas, the stimuli in this study defined hypotheses in an abstract similarity space over geometric shapes. Participants could request labels for either category members or non-members. While both request types were used in all conditions, most often evenly, the proportion of participants showing a preference for one type of request was strongly impacted by the information value of that request type. A small tendency to prefer requests from the designated target category was also observed.

Keywords: hypothesis testing; positive test bias; sparsity; information sensitivity;

Introduction

Very few people ever become professional chicken-sexers, but most of us can identify with the *kind* of problem they face. Collections of ambiguous targets that need to be categorized are common. Opal miners, worlds away from poultry-breeding, face a similar problem sorting through piles of nearly identical rocks for the few that, if polished, may reveal a gemstone shine. Consider the problem faced by a novice in each of these fields trying to become expert via training examples from some authoritative test, perhaps an expert teacher. What kinds of examples should they seek in order to become an expert themselves as quickly as possible? At first glance, different problem domains require different strategies. A natural strategy for the chicken-sexer might be to request an even number of examples of both male and female chicks. The beginner opal miner on the other hand, is probably only interested in seeing examples of opals in the raw. The central claim of this paper is that these differing tendencies can be described under the same normative framework, and reflect a rational sensitivity to features of the problem domain: in this case that chicks of either sex are nearly equally likely, while opals are rare among rocks.

Since they are tasks that involve some learner control over what information will be received, novice opal mining and chicken sexing can be considered examples of active learning (Settles, 2009; Gureckis & Markant, 2012). The psychological literature has considered a number of different normative standards against which human behavior in such tasks can be assessed (Nelson, 2005). The traditional approach comes from the philosophy of science and treats falsificationism as the normative standard. According to this view, learners should conduct tests designed to falsify their current hypothesis, on the grounds that confirming evidence is always open to alternative explanations, but counterexamples always definitively rule out a hypothesis (Popper, 1959). Strict falsification is rarely followed by people faced with hypothesis-testing tasks (Wason, 1960, 1968). Instead, a tendency to propose tests consistent with a working hypothesis has been replicated in a wide range of tasks and contexts (Nickerson, 1998).

Although originally considered to be an irrational bias, people's tendency to seek positive tests may have a solid statistical basis. Tests consistent with a currently preferred candidate hypothesis can falsify the hypothesis in situations where the true hypothesis is not a superset of the candidate one (Klayman & Ha, 1987). In choosing whether to probe from within the scope of a candidate hypothesis or outside it, the learner must consider the base rate probability that a member of the domain under consideration is also a member of the target set, the proportion of domain members that are covered by the candidate hypothesis, and (estimated) positive and negative error rates under the candidate hypothesis (Klayman & Ha, 1987). When the target set is a relatively small subset of the whole domain, and its size is approximately known, positive testing can maximize the chance of falsification.

Expanding on the work of Klayman and Ha (1987), recent studies have tended to assess the quality of a particular query in statistical terms. A good query might be one that minimizes the expected probability of error on a randomly selected domain member after the seeing the results of the test (expected probability gain), or returns the most information about the identity of the true hypothesis (expected information gain). Although differing in their predictions under some circumstances, these measures share the idea that a hypothesis test is a kind of gamble with uncertain rewards in terms of evidential value (Poletiek & Berndsen, 2000). Unlike strict falsification, a strategy of maximizing expected probability gain has been shown to account well for human responses in simplified hypothesis testing tasks (Nelson, McKenzie, Cottrell, & Sejnowski, 2010).

This recent line of work opens up an important question: does people's preference for positive evidence genuinely reflect a sensitivity to its informational value, or is it a cognitive bias that just happens to produce good results in some tasks? It is this question that we consider in this paper.

Hypothesis sparsity and information search

Theoretical results showing the value of positive evidence do not imply that the positive test strategy is universally the best approach (Klayman & Ha, 1987; Austerweil & Griffiths, 2011; Navarro & Perfors, 2011). Rather, they imply that it works when the possible hypotheses are *sparse*. If hypotheses are thought of as indicating meaningful subsets of some larger domain, the sparsity of a hypothesis refers to the proportion of all members of a domain that are selected. Sparse hypotheses include fewer than half of the members of the relevant domain (Navarro & Perfors, 2011). For example, in the domain LIVING SPECIES the category DOGS is sparse, while AEROBIC ORGANISM is a not sparse, since most living things are not dogs, but do metabolise oxygen. Sparsity can vary in degree: while DOGS and POODLES are both sparse categories in the domain of living things, the category POODLES is more sparse. Arbitrarily complex hypotheses can still be described in terms of a set of members, for example when describing a linguistic hypothesis in terms of the set of sentences the hypothesis considers grammatical.

Sparsity is one factor impacting the utility of the positive test strategy (Klayman & Ha, 1987; Navarro & Perfors, 2011). Where the target hypothesis is sparse, the expected information value of negative tests is reduced. As figure 1 shows, in this case the probability of a negative test producing a disconfirming positive result is low, regardless of the accuracy of the learner's working hypothesis. Conversely, if the target hypothesis is not sparse, the value of positive testing falls, because most positive tests return an expected and therefore uninformative positive result. This yields a natural prediction: manipulating the sparsity of the learner's hypotheses should impact the degree to which the learner favours positive tests. If the hypotheses considered are not sparse, an optimal learner should show a negative test bias.

There is evidence suggesting this pattern is observed empirically. For instance, Hendrickson, Navarro, and Perfors (in preparation) had participants play a modified version of the game "battleships". Hypotheses corresponded to a possible configuration of ships, each of which covered an area in a two dimensional space. The size of the ships was varied across experimental conditions, changing the sparsity of hypotheses. The predicted effect was observed: when hypotheses were not sparse, people shifted away from positive tests and towards negative ones.

One potential problem with this result is that the task was highly visual rather than conceptual in nature. In this case, the coverage of the possible domain by a given hypothesis was also literally the coverage of a 2D field. It is possible that the estimation of relative likelihoods apparently considered by participants was driven by an estimate of relative area particular to the visual system. In order to generalize to other active learning tasks, it must be established that the idea of coverage in a domain extends from literal physical spaces to abstract conceptual spaces. It is also possible that the extra difficulty associated with such abstract spaces makes reliance on simple heuristics such as a positive testing bias more attractive (Cherubini, Rusconi, Russo, Di Bari, & Sacchi, 2010).

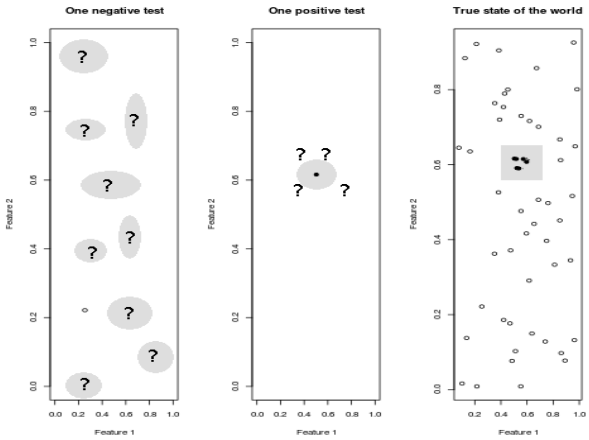

Figure 1: A toy world showing the information value of positive and negative requests with a sparse category. Objects in this world vary on two features, x and y, and some representative plausible category boundaries are drawn in grey. In the first panel, the learner has a single example outside the target category. This example is uninformative as it is consistent with many category structures. In contrast, the second panel shows a learner who has a single example of a category member. A large number of the hypotheses in the first panel are ruled out, leaving only those which include the observed example. The third panel shows the critical properties of the world generating the examples for these learners: the target category is in fact small and coherent. If it were not, the advantages of positive testing would be reduced or even reversed (Navarro & Perfors, 2011).

The current study aims to extend the results in Hendrickson et al. (in preparation), to see if the same effect can be observed with stimuli drawn from a more abstract stimulus space. The experiment took the form of a sorting task asking participants to learn a category boundary in a stimulus space consisting of simple geometric shapes varying on three feature-dimensions, described below. Participants were able to request labels for a randomly selected positive example of the target category or a negative non-target example. People were aware of what proportion of all stimuli belonged to the target category, and between-subjects manipulations of this proportion showed both a sensitivity to the information value of each type of request and a small preference for requesting positive examples.

Method

Participants

367 adults were recruited via Amazon Mechanical Turk. Of these, 301 completed the task, and 121 were excluded from further analysis for either failing to make any label requests at all (85 participants), making more than 60 requests (nine participants), or failing to sort labelled examples into the category indicated by the label, revealing a lack of effort or a misunderstanding of the task (36 participants). Nine participants were excluded for a combination of these reasons.The remaining 180 participants completed one practice run and two real runs of the sorting task, with between 104 and 131

completed sorts recorded in each of three sparsity conditions. These conditions set the proportion of stimuli belonging to the target category at 25% SPARSE, 50% EVEN, or 75% NON-SPARSE. Note that the only difference between the SPARSE and NON-SPARSE conditions is one of framing, while both differ from the even condition in terms of the relative information value of requests.

Ages ranged from 19 to 67 (mean: 34.4) and 45.0% were female. 117 of the final participants were from the United States and 52 were from India. Those remaining were from eight other countries in Africa, North and South America, Europe, and Asia. All participants were paid $0.60US for the 15 minute experiment.

Procedure

The cover story for the study described a fictitious company interested in harvesting a new substance called selenoid from plankton. Participants were told selenoid-rich plankton were desirable for harvesting, and were given the percentage of all plankton expected to be selenoid-rich, either 25%, 50%, or 75% depending on the experimental condition. In each trial, people were presented with two bins, each containing a random selection of half the possible plankton examples, as shown in figure 2. Buttons below each bin allowed participants to request a label for either a selenoid-rich or a selenoid-poor plankton, which appeared as a persistent colored highlight around a randomly selected example of the requested type after a two-second delay. Plankton could be swapped between bins by clicking on them, and participants were asked to click a submit button after they had sorted each plankton into the correct bin.

Once a sort was submitted, the true selenoid status for each plankton was revealed and a score displayed, calculated as 10 points for each plankton correctly sorted and -10 for each incorrectly sorted. An inference-efficiency score defined as total score divided by number of requests made was also displayed. The maximum score under this scheme was 640. The expected score due to chance was zero in the EVEN condition and 160 in the SPARSE and NON-SPARSE conditions.

The experiment began with all participants answering a series of multiple-choice questions to make sure they had read and understood the instructions. The main task was then presented three times, the first of which was labelled as a practice trial and required participants to try all the available actions and submit a sort that respected the known proportion of plankton in each category and any visible labels. Data from this trial was not analysed. Stimuli were colored either red, blue, or green (order randomized) across the three trials to emphasise their distinctness, with color variation between stimuli within a single trial based on four evenly spaced points on the 255-value RGB scale for that color.

Stimuli

The stimuli were geometric shapes consisting of a ring and a number of radial arms. They varied in color intensity, size of the ring, and number of arms, with four levels in each dimension giving 64 combinations of feature values.

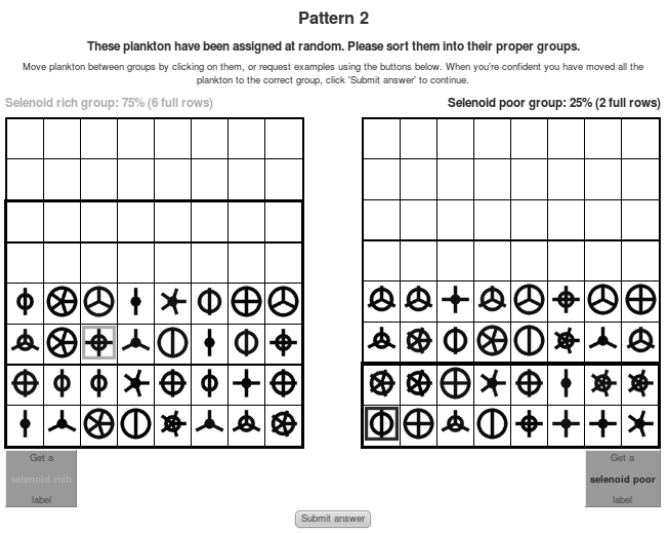

Figure 2: Presentation of the sorting task. Information available at all times included all possible plankton examples, the proportion of plankton belonging to each group, and the request types available. Labels, if requested, appeared as a persistent colored border around a randomly selected example of the appropriate type. An initial configuration is shown here, but two requests have been made, one of each type.

Figure 3: 64 different stimuli were used, corresponding to all unique combinations of four possible values on three dimensions. These were color, ring size, and number of arms, shown here increasing from left to right.

The true selenoid status of the plankton in a given trial was determined by a threshold rule on one dimension of variation, randomly selected under the constraint that rules could not repeat across the three trials presented to any one participant. The location of this threshold was determined by sparsity condition, which varied between participants. In the SPARSE and NON-SPARSE conditions, members of the minority group shared one extreme value on one type of feature. In the EVEN condition, members of the same group shared one of two adjacent values in the discriminating feature. For example, a participant in the SPARSE condition might view a practice trial using red plankton in which selenoid-rich plankton had four arms and selenoid-poor plankton one, two or three, then view a trial using blue plankton where only plankton with the largest circles were rich, and finally a trial using green plankton where only the darkest shade of green were rich. Although repetition of another rule using number of arms could not have been presented to this participant after the first trial, the order of rules and whether the thresholds were high or low were completely randomized.

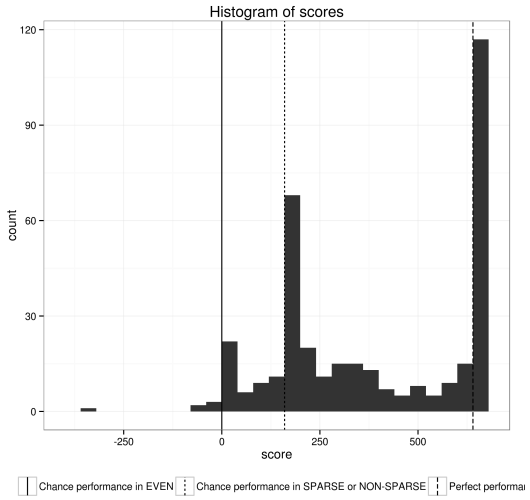

Figure 4: The modal score in all conditions was near perfect performance, although many participants sorted with chance performance, shown as peaks in the histogram at zero in the EVEN condition and 160 in the SPARSE and NON-SPARSE conditions

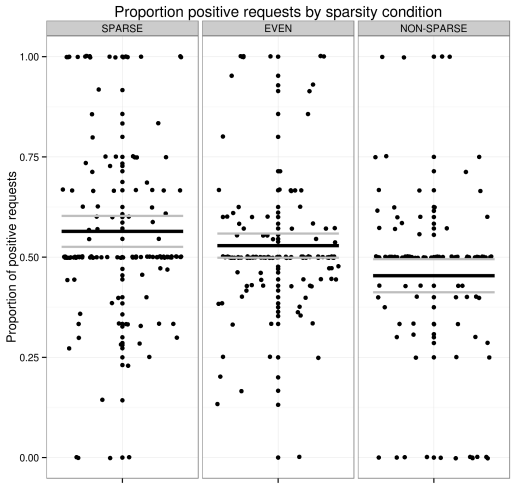

Figure 5: Proportion of positive requests appear to cluster at values zero, 0.5, and one. Means and 95% confidence intervals are plotted as dark and light horizontal lines respectively. The observed differences between means may be driven by participants preferentially choosing between a small number of distinct request strategies.

Results

The comparisons of interest between conditions required that participants be engaged with the task. The average score across participants was 368 of a maximum 640, corresponding to 50.4 of 64 plankton correctly sorted.As figure 4 shows, score distributions were bimodal in each condition, with one peak at the expected score due to chance (zero for EVEN and 160 for SPARSE and NON-SPARSE) and the other peak at perfect performance.

While 27% of trials scored at or below chance, many people were highly successful: on 18% of all trials, people required fewer than six labels to sort at least 93% correctly. The mean number of swapping actions (36.0) was close to the expected required number of swaps to correct a random sort to an ordered one (32), indicating that participants meeting the inclusion criteria above understood and engaged with the task. Scores and number of label requests were not significantly different across the first and second non-practice trials (two sample Kolmogorov-Smirnov test, $D = .0487$, and $D = .0339$ respectively, $p > .98$).

Where positive testing bias predicts a preference for requests labelling the selenoid-rich plankton category regardless of the population proportions, sensitivity to the information value of requests implies a preference for requesting labels from the minority classification if this is possible. We distinguish among these hypotheses by comparing the proportion of positive requests in each trial across conditions.

As figure 5 shows, the proportion of positive requests differed across sparsity conditions ($F(2,359) = 9.581$, $p < 0.001$). A post-hoc Tukey Test showed that the proportion of positive requests was significantly higher in the SPARSE than NON-SPARSE and in EVEN than NON-SPARSE. Potential nuisance variables trial color, trial number, and left/right order of presentation were not found to have a significant effect (did not improve model AIC).

As Figure 5 shows, the differences in means do not appear to reflect a large-scale shift in requests for all individuals from condition to condition. Rather, in all conditions, responses tended to cluster at the values of one (all positive), zero (all negative), and 0.5 (even). The proportion of people choosing each strategy is what appears to shift between conditions. We can test whether this is truly occurring by categorizing responses by request strategy, as in Figure 6. A participant was classified as using an *Even* request strategy if their proportion of positive requests fell between 0.45 and 0.55, with *Prefers positive* and *Prefers negative* responses falling above and below these values.

All strategies were followed by some participants in all conditions, but the proportion of participants selecting each strategy differed significantly between conditions ($\chi^2_2 = 13.16, p < .01$). The *Even* request strategy balancing positive and negative requests was always most popular, although the proportion of people preferring positive tests was higher in the SPARSE condition than in the NON-SPARSE condition. The inverse pattern was observed for negative tests, indicating a preference for whichever type of test corresponded to the minority classification in the whole population. Such a preference is consistent with the greater information value of minority requests. As Figure 6 shows, this preference is evident even when the minority group is framed as a negative non-target group.

Since the only difference between the SPARSE and NON-SPARSE conditions is one of framing, if request choice is driven by expected information gain, request preferences should be simply reversed between these two conditions. However, figure 6 shows a slight asymmetry whereby positive preferential strategies are more often chosen. Collapsing across conditions, 0.48 of all information requests recorded

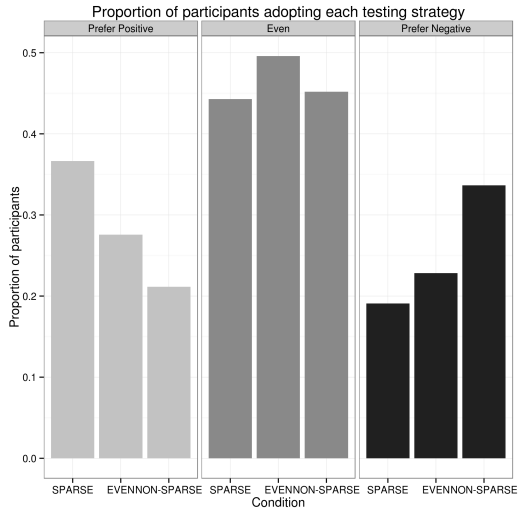

Figure 6: Proportion of participants using each testing strategy in the three conditions. Requesting equal amounts of information from both possible categorizations was popular in all conditions, however when population proportions were unequal, a greater proportion of participants began to prefer requests from the minority group. This preference was somewhat asymmetrical, with people more readily switching to preferring positive requests

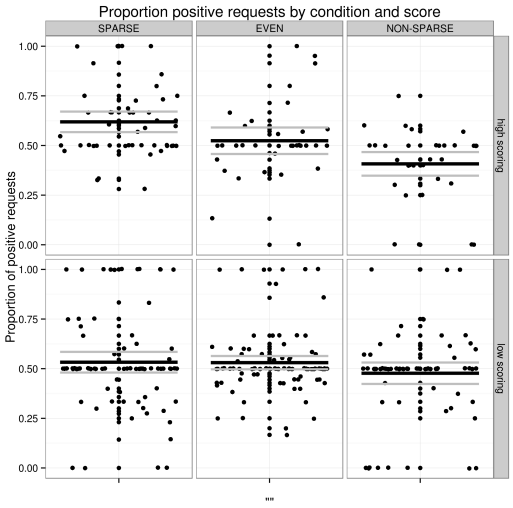

Figure 7: Differences in mean proportion positive requests significant in the higher scoring group, and not significant, although in the same direction, in the lower scoring group.

in this study were for selenoid-poor labels (1192 of 2482 total requests), and 0.52 were for selenoid-rich labels (1290 of 2482 total requests), a small but significant difference in favour of positive requests despite the symmetry of the task structure ($\chi^2_1 = 7.74, p < .01$).

To examine whether people's apparent sensitivity to sparsity was in fact driven by a subset of participants, we partitioned trials by score into high-scoring (score over 600, n=117), and lower-scoring trials (score less than or equal to 600, n=245). As figure 7 illustrates, both high-scoring trials and low-scoring trials showed the same clustering pattern at zero, 0.5, and one, with 0.5 the most popular strategy. Also the tendency to switch to preferring labels of the minority categorization is significant only in the higher scoring group (high scoring: $F(2,114) = 13.04$, $p < .01$, low scoring: $F(2,242) = 1.63$, $p > .1$).

Discussion

The results show that people adjust their sampling strategies in response to the sparsity of the hypothesis testing task at hand, supporting an information sensitive account of natural hypothesis testing (Navarro & Perfors, 2011). This sensitivity to the relative size of the target category in the stimulus space obtains even though it is an abstract space defined over the similarity of a set of geometrical shapes. The observed behavior is inconsistent with a strong bias towards positive testing, which predicts a difference between the symmetrical SPARSE and NON-SPARSE conditions due to the change of frame. No strong difference was observed, although positive tests were more popular than negative tests overall.

Individual participants tended to use either balanced requests or requests of a single type, but the popularity of these strategies varied across conditions. The *Even* strategy was always most popular. This may reflect a popular explorative heuristic that motivates people to use all available request types, or possibly simply a lack of engagement with the task. This second alternative is somewhat supported by the observation that preferential strategies were used more among the highest scoring participants. Despite the overall popularity of the *Even* strategy, the attractiveness of preferential strategies scaled as predicted with changes in the true information value of requests. For this to be the case, the value of requests must have been estimated (not necessarily explicitly) by participants from information available about category sparsity. Such information is often available as domain knowledge in real-world category learning tasks, as it is in the opal mining and chicken sexing problems. It can also be estimated, even when the universe of all possible examples is large, for example by estimating a base rate probability from an observed frequency or through capture-recapture estimation techniques (also not necessarily explicit) when repeatedly encountering novel and familiar examples of a category.

The behavior observed in the context of the plankton-sorting task suggests that to the extent that category sparsity information is available, it could be expected to impact the perceived attractiveness of different types of information request, positive or negative. That said, although consistent with a degree of sensitivity to the information value of requests, these data show a number of ways in which people's requests deviate from information-sensitive treatments of the task.

Participant showed some positive test bias, favouring the target rich category asymmetrically despite the symmetry of the sparsity manipulation. It is unclear if this is due to a form of matching (Evans, 1998) on the target most prominent in the instructions, or an expectation that the conditions that favour positive testing are generally ubiquitous, although in this ar-

tificial case they are not.

The clustering of positive-test proportions at zero, 0.5, and one in all conditions also suggests a kind of heuristic approach, albeit a heuristic that is to some extent context sensitive. It is unclear from these results if this clustering is reflective of granularity in the perception of information utility, granularity in responding after accurate perception of information utility, or simply an artifact of the fact that participants were limited to two different request types, which to some extent naturally emphasises these values, especially for small numbers of requests.

A number of features of the presentation are also open to question. All possible plankton shapes were visible to participants at all times, a situation unlikely with natural categories, but one which might influence the use of sparsity information, since estimating the proportion of stimuli indexed by a hypothesis in a given domain requires an estimate of the boundaries of that domain. Similarly, the density of examples was even across the stimulus space, with one example of each kind of plankton, a condition which need not hold in general. The true category rules were also highly restricted, in that they were all thresholds on a single dimension, binary, and strictly complementary. Natural categories are often non-binary, nested or otherwise overlapping, and often involve multiple dimensions.

Further work is required to explore how people weigh up the value of information-seeking actions under these more complex conditions. The results presented here suggest that some such consideration of information utility is an essential component of any complete description of natural active concept learning. Where hypothesis sparsity information is available, as it is in many natural domains, people's information-seeking behavior is informed by the structure of the domain, even when that structure is abstract and conceptual in nature.

Acknowledgements

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. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*(3), 499–526.

Cherubini, P., Rusconi, P., Russo, S., Di Bari, S., & Sacchi, S. (2010). Preferences for different questions when testing hypotheses in an abstract task: Positivity does play a role, asymmetry does not. *Acta psychologica*, *134*(2), 162–174.

Evans, J. S. B. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking & Reasoning*, *4*(1), 45–110.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning a cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Hendrickson, A. T., Navarro, D. J., & Perfors, A. F. (in preparation). 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.

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

Nelson, J. D. (2005). Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4), 979.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters information acquisition optimizes probability gain. *Psychological science*, *21*(7), 960–969.

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

Poletiek, F. H., & Berndsen, M. (2000). Hypothesis testing as risk behaviour with regard to beliefs. *Journal of Behavioral Decision Making*, *13*(1), 107–123.

Popper, K. R. (1959). The logic of scientific discovery. *London: Hutchinson*, *1*.

Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report No. 1648). University of Wisconsin–Madison.

Wason, P. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, *12*(3), 129–140.

Wason, P. (1968). On the failure to eliminate hypotheses: A second look. *Thinking and reasoning*, 165–174.