# Sensitivity to hypothesis sparsity in a category discrimination task

**Steven Langsford** (steven.langsford@adelaide.edu.au)
**Drew 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

People's sensitivity to expected information value when choosing between two different types of information request was examined in a simple category-learning task. Previous work has shown that the true 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. However the degree to which people are sensitive to expected information value is a subject of ongoing research. This study examined a binary sorting task where sparsity differed across conditions. In contrast to previous work using visual areas to represent the coverage of a domain by a particular hypothesis, the stimuli used in this study defined hypotheses in an abstract similarity space over geometric shapes. Participants were able to request a label for examples of either category members or non-members. The true value of these requests varied depending on the proportion of all stimuli which belonged to the target category, which was known by participants. 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. These results are discussed in the context of previous work showing information sensitivity and positive test biases in hypothesis testing tasks.

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

## Introduction

As you read this, countless toddlers are learning their first language by producing sentences that may or may not be grammatical, would-be master chefs are adding things to cakes that may or may not be a good idea, and novice basketball players are turning to their friends or coaches and asking "So was *that* a foul? Come on, surely that was a foul?". Diverse as these activities are, they are all examples of active learning, tasks where the learner has some control over what information will be recieved. The learner may be able to produce an example of a putative category for validation, as for the toddlers and would-be chefs, or requesting a label for an opportunistically received exemplar, as in the case of the unfortunate basketball player, but in either case, the question arises of how to do this efficiently (Settles, 2009; Gureckis & Markant, 2012). Toddlers cannot possibly test all possible combinations of words, however much they might like to, and athletes cannot challenge every play. Exactly which examples to test though, is a non-trivial problem, and the optimal solution varies with properties of the problem domain, with good baking strategies potentially not generalizing well to the basketball court.

So what are these good strategies? The psychological literature has considered a number of different normative standards against which human behavior 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). Although widely accepted as a scientific norm, 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. One important observation, due to Klayman and Ha (1987), is that tests consistent with a currently preferred candidate hypothesis can in fact falsify the hypothesis in situations where the true hypothesis is not a superset of the candidate one. 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 is a defensible strategy in terms of maximizing the chance of falsification of a candidate hypothesis.

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 – in information theoretic terms – about the identity of the true hypothesis (expected information gain). Although differing in their predictions under some circumstances, what these measures share is the idea that a hypothesis test is a kind of risky gamble that returns uncertain rewards in terms of evidentiary 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

To determine whether or not people are considering the informational value of different types of evidence, it is useful to recognize that the theoretical results showing the value of positive evidence (Klayman & Ha, 1987; Austerweil & Griffiths, 2011; Navarro & Perfors, 2011) do not imply that the positive test strategy is universally the best approach. Rather, they imply that it works when the possible hypotheses are *sparse*. The sparsity of a hypothesis refers to the proportion of all members of a domain that are indexed by the hypothesis in question. Sparse hypotheses index 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 non-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.

Sparsity is one factor impacting the utility of the positive test strategy (Klayman & Ha, 1987; Navarro & Perfors, 2011). Where the target hypothesis is very sparse, the expected information value of negative tests is greatly reduced, because the probability of a negative test producing a valuable disconfirming positive result is very low. Even when the candidate hypothesis is completely misplaced, testing outside it is unlikely to land on such a small critical area, and if it is approximately correct, even this small chance is reduced. Conversely, if the target hypothesis is non-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 produce a corresponding effect on the strength of the positive test bias. In fact, if hypotheses tend to be dense (i.e., are consistent with a majority of possible tests), an optimal learner should show a negative test bias. (See Figure 1 )

There is some evidence to suggest that this pattern is observed empirically. For instance, Hendrickson (in preparation) had participants play a modified version of the game "battleships". In this task, hypotheses corresponded to a possible configuration of "ships", each of which covered an area in a two dimensional space. In this task, all hypotheses have the same sparsity (because the ships always cover the same area). Across experimental conditions, the size of the ships was varied, leading to a change in the sparsity of hypotheses. The predicted effect was observed: when hypotheses became dense, people tended to shift away from positive tests and towards negative ones.

One potential problem with this result is that the task was highly visual, not "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, making it 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).

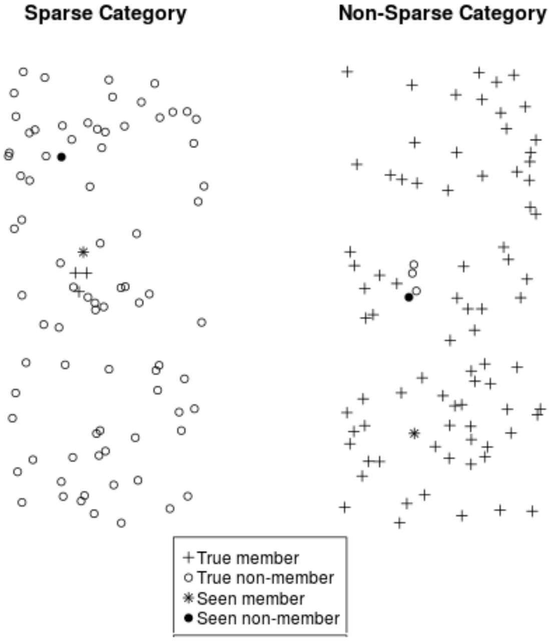


Figure 1: Two simulated categories for objects varying on two dimensions. In the first case, where the category is sparse, knowing the location of a single category member is highly informative about the location of the others, and knowing the location of a single non-member is nearly totally useless. In the non-sparse case, this situation is reversed. Positive testing is optimal when targets are sparse and coherent.(Navarro & Perfors, 2011)

The current study aims to extend the results in Hendrickson (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. Participants 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 (9 participants), or failing to sort labelled examples correctly (36 participants). Nine participants were excluded for a combination of these reasons.The remaining 180 participants contributed 360 trials, with between 104 and 131 trials falling in each of three sparsity conditions. These conditions set the proportion of stimuli belonging to the target category at 25%, 50%, or 75%, hereafter SPARSE, EVEN, and 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), 45.0% were female. 117 of the final participants were from the United States and 52 were from India. Those remaining were from 8 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, participants were presented with two 'bins', each containing a random selection of half the possible plankton examples (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 coloured 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 zero in the EVEN condition and 160 in the SPARSE and NON-SPARSE conditions.

All participants answered 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 coloured either red, blue, or green (order randomized) across the three trials to emphasise their distinctness, with colour variation between stimuli within a single trial based on four evenly spaced points on the 255-value RGB scale for that colour.

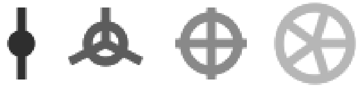


Figure 3: 64 different stimuli were used, corresponding to all unique combinations of four possible values on three dimensions. These were colour, ring size, and number of arms, shown here increasing from left to right.

## Stimuli

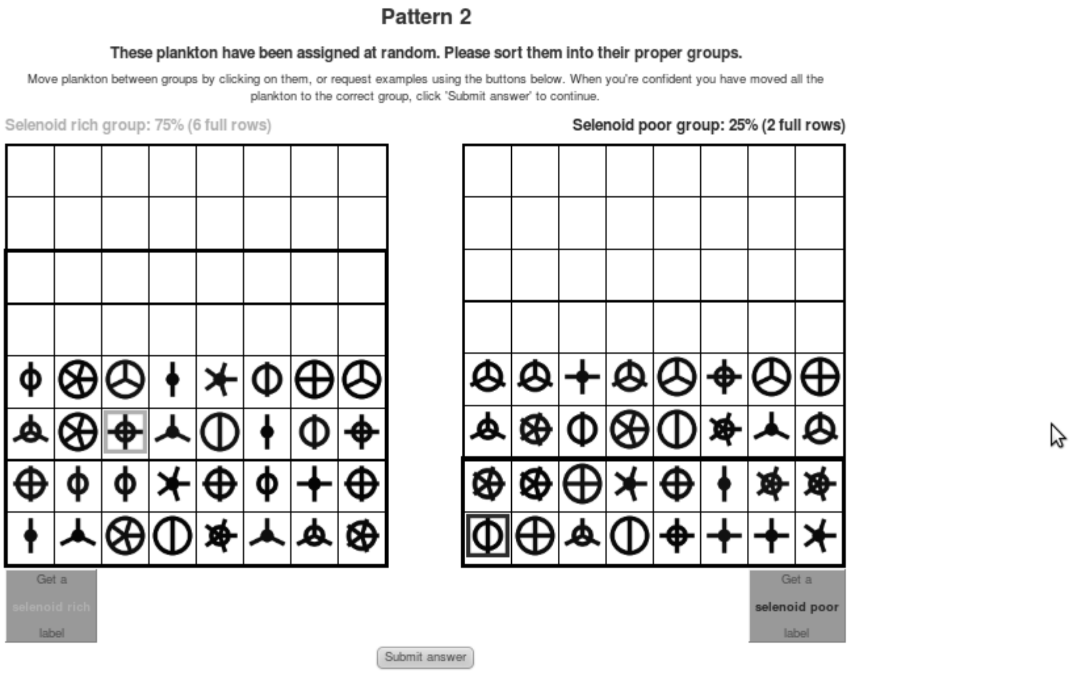
The stimuli were geometric shapes consisting of a ring and a number of radial arms. They varied in colour intensity, size of the ring, and number of arms, with four levels in each dimension giving 64 combinations of feature values.

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 'rich' plankton had four arms and '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.

# 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. Score distributions were bimodal in each condition, with one peak at the expected score due to chance (0 for EVEN and 160 for SPARSE and NON-SPARSE) and the other peak at perfect performance (Figure 4).

While 27% of trials scored at or below chance, many people were highly successful, defined as able to sort 60 or more plankton correctly on the basis of fewer than 6 labels (18% of all trials). 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 par-



Pattern 2

These plankton have been assigned at random. Please sort them into their proper groups.

Move plankton between groups by clicking on them, or request examples using the buttons below. When you're confident you have moved all the plankton to the correct group, click 'Submit answer' to continue.

Selenoid rich group: 75% (6 full rows)

Selenoid poor group: 25% (2 full rows)

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 coloured 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.

ticipants 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 K-S test, $p > .98$).

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. The proportion of positive requests in each trial was compared across conditions.

The proportion of positive requests differed across sparsity conditions (F(2,359) = 9.581, p¡0.001, see Figure 5). Specifically, the proportion of positive requests was significantly higher in the SPARSE than NON-SPARSE (95% HSD between .17 and .04, $p < .001$), and in EVEN than NON-SPARSE(95% HSD between .14 and .01, $p < .02$). Potential nuisance variables trial colour, trial number, and left/right order of presentation were not found to have a significant effect (did not improve model AIC).

The differences in means appear to be due to a qualitative shift between distinct request strategies, with participants' responses tending to cluster at the special values of 1, 0.5, and 0 positive requests (see Figure 5).

This clustering of request proportions motivates a categorization of responses by request strategy, 'Prefers positive', 'Even' and 'Prefers negative'. An 'Even' request strategy was defined as a proportion of positive requests falling between 0.45 and 0.55, with 'Prefers positive' and 'Prefers negative' responses falling above and below these values (Figure 6).

All strategies were followed by some participants in all conditions, and the 'Even' request strategy balancing positive and negative requests was always the most popular, never lower than 44% of participants. The population percentage of 'rich' examples did impact on the attractiveness of each testing strategy: the proportion of people preferring positive tests fell from 37% in the SPARSE condition to 21% in the NON-SPARSE condition, while the proportion of people preferring negative tests rose from 19% to 33%, indicating a preference for whichever type of test corresponded to the minority classification in the whole population. When the plankton population was EVEN, the 'Even' testing strategy was more popular than otherwise at 50% of all participants, and a preference for positive tests was more common that a preference for negative tests, at 27% and 23% of participants respectively.

In the SPARSE and NON-SPARSE conditions, there exists a highly efficient strategy under the threshold rules used where only two examples of the minority category need be requested in order to give a 75% chance of uniquely determining the rule used. To examine whether participant's apparent sensitivity to sparsity was in fact driven by a subset of participants deducing the role of sparsity in this particular task and

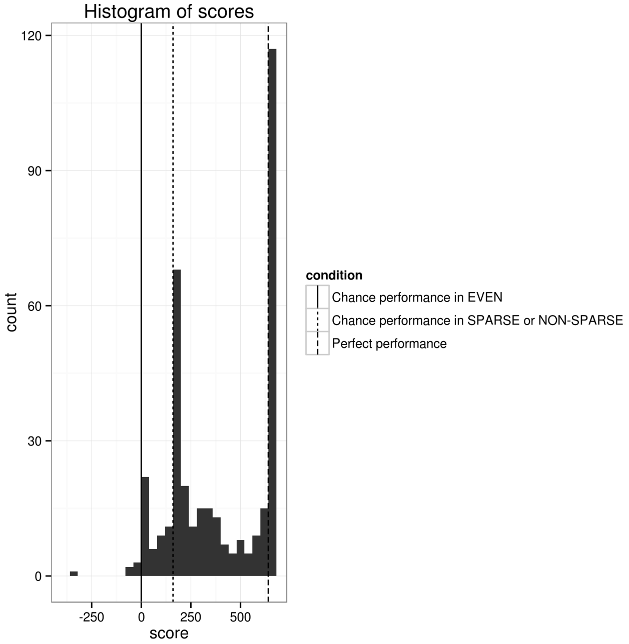


Figure 4: The modal score in all conditions was near perfect performance, although many participants sorted with chance performance, shown as peaks in the histogram at 0 in the EVEN condition and 160 in the SPARSE and NON-SPARSE conditions

producing atypical behaviour, trials were partitioned by score into high-scoring 'solved' trials (score over 600, n=117), and lower-scoring trials (score less than or equal to 600, n=245). Both high-scoring trials and low-scoring trials showed the same clustering pattern at 0,.5,and 1, with .5 the most popular strategy. The qualitative pattern of these data were unchanged under this subdivision: both show an increased tendency to switch to preferring labels of the minority categorization, and a slight asymmetry in favour of positive tests.

To determine if the asymmetry in favour of positive tests was statisically significant, a two-sample test for equality of proportions was carried out on corresponding strategy/condition pairs. That is, the proportion of participants using a prefers-positive strategy when positive examples were in the majority (22 of 104) was compared with the proportion of participants using a prefers-negative strategy when negative examples were in the majority (25 of 131) and so on. An alpha of .983 was adopted for each test to maintain a significance level of .95 across three comparisons. Although all differences were in the expected direction, only one was significant by this measure. The proportion of participants switching to a prefers-positive strategy when positive examples were in the minority was significantly greater than the proportion switching to a prefers-negative strategy when negative examples were in the minority ($p < .004$). Collapsing across conditions, .48 of all information requests recorded in this study were for 'selenoid-poor' labels (1192 of 2482 total requests), and .52 were for 'selenoid-rich' labels (1290 of 2482 total requests), a small but significant difference ($p < .01$).

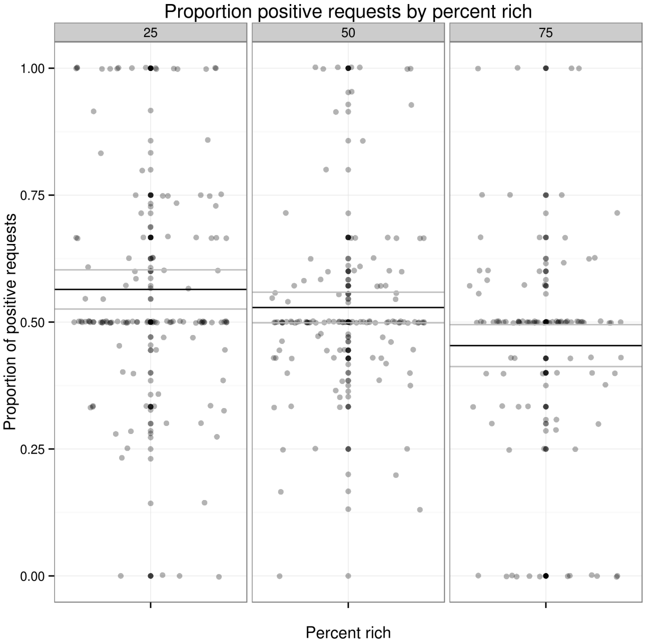


Figure 5: Proportion of positive requests appear to cluster at values 0, 0.5, and 1. Means and 95% confidence intervals are plotted as dark and light horizontal lines respectively. The observed differences between means may be driven by participants preferentially choosing between a small number of distinct request strategies.

## Discussion

The results show people adjusting 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 this is an abstract space defined over the similarity of a set of geometrical shapes. The effect appears more pronounced among participants who achieved high scores, but the predicted sparsity-based sampling differences are also apparent among participants with less-than-perfect scores, making it unlikely that the pattern observed is due to a subset of participants deducing an optimal strategy for this particular task.

Individual participants reflected sensitivity to the information value of the different request types in a coarse-grained way, gravitating to either balanced requests or requests of a single type. However the attractiveness of a particular strategy, as reflected by the probability of a participant in a given condition choosing that strategy, 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 the information available about the proportion of all examples falling in the target category. Such information is often available or can be estimated in real-world category learning tasks, even when the universe of all possible examples is large, for example by estimating a base rate probability from an observed frequency, or through more sophisticated capture-recapture

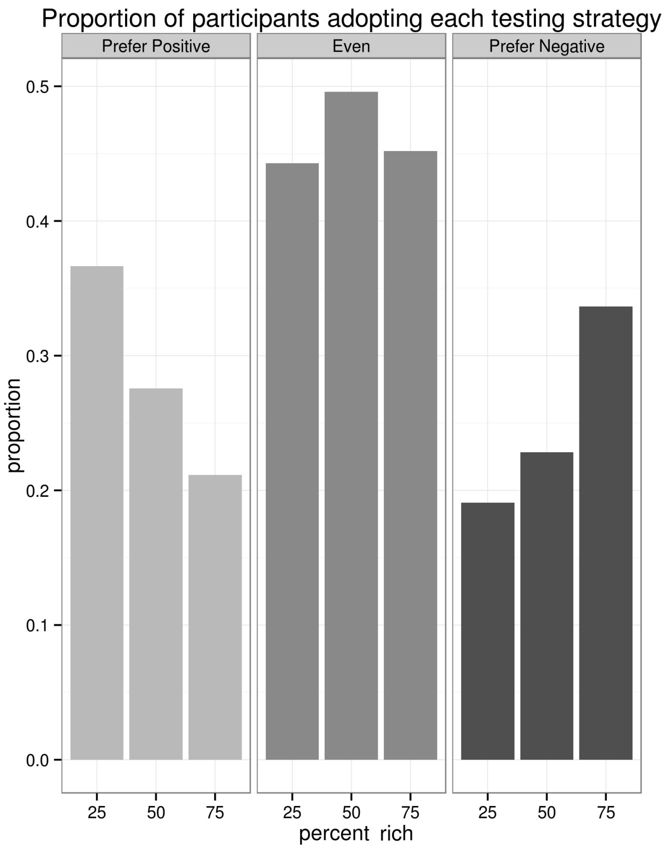


Figure 6: Proportion of participants using each testing strategy in the three conditions. Requesting equal amounts of information from both possible categorizations was popular in all conditions, however when population proportions were unequal, a greater proportion of participants begin to prefer requests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

estimation techniques (also not necessarily explicit) when repeatedly encountering novel and familiar examples of a category.The behaviour 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, ie. preferences for positive or negative testing.

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-utility treatments of the task.

A positive testing bias is suggested, with participants favouring the target 'rich' category asymmetrically despite the symmetry of the task under 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 artificial case they are not.

The clustering of positive-test proportions at the special values 0, .5, and 1 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 request 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.

These conclusions are also subject to a number of limitations. The dropout rate was high (656 views of the instructions resulting in 365 attempts, 301 completions, and 180 participants meeting inclusion criteria), raising the possibility that this self-selected sample is unrepresentative. 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 both heuristic accounts and expected information-utility accounts will be needed.

## Acknowledgements

This research was supported by...

## 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, D. (in preparation). ¡¡working title¿¿ battleships.

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.