Sensitivity to hypothesis size during information search

Andrew T. Hendrickson
School of Psychology
University of Adelaide

Danielle J. Navarro
School of Psychology
University of Adelaide

Amy Perfors
School of Psychology
University of Adelaide

Abstract

It is well known that people attempting to perform hypothesis testing show a positive test bias, preferring to request evidence that is consistent (rather than inconsistent) with their current hypothesis. Rather than viewing this as an irrational bias, information theoretic accounts of hypothesis testing have argued that selecting tests likely to produce positive evidence is adaptive when most hypotheses are small (i.e., true of few entities in the world) and thus respond positively to very few queries. These accounts make the prediction that as hypotheses get larger, the relative utility of positive evidence will decrease; when hypotheses are large enough, negative evidence will become more useful than positive evidence. We test if people are sensitive to this change in utility with an experiment inspired by the game "Battleship" in which people attempt to discover the correct arrangement of ships by asking for positive or negative evidence. As predicted, as hypotheses become larger people request less positive evidence, and when hypotheses are large requests for negative evidence are more likely than requests for positive evidence. Implications for the nature of the positive test bias are discussed.

AFP received salary support from ARC grant DE120102378 and DJN from ARC grant FT110100431. Research costs and salary support for ATH were funded through ARC grant DP110104949. Preliminary version of this work was presented at the 46th Annual Meeting of the Society of Mathematical Psychology.

Introduction

In many real life situations, asking the right questions is central to successful learning. Scientists design experiments to probe the state of the world, police interview witnesses to solve crimes, children ask parents questions to guide their learning, and so on. In all these cases, there is a learner (scientist, police officer, child) who poses queries to a teacher or the world (nature, witness, parent), and uses the feedback from this source to decide among rival hypotheses. What queries should the learner make in order to test hypotheses? What strategy do people typically use? And what factors influence the strategies people should and do employ? In the current paper we focus on the latter question, in particular the importance of the ratio of positive to negative evidence of the hypotheses to be tested. We show that people are highly sensitive to this ratio, the "size" of the possible hypotheses, when determining what sort of information to ask for.

The psychological literature on how people should test hypotheses has borrowed heavily from the logic of scientific discovery. The standard version of the scientific method taught in every science classroom places the emphasis on falsification of hypotheses: the idea that the learner should pick a candidate hypothesis and conduct an experiment that has the power to eliminate that hypothesis if it is in fact false (Lakatos, 1970; Platt, 1964; Popper, 1934/1959). The overwhelming body of evidence shows that intuitive hypothesis testing by human learners is generally not falsificationist in character. That is, in a huge variety of situations, people do *not* select queries that are well-suited for falsifying their current hypothesis. This pattern has been observed in concept learning (Bruner, Goodnow, & George, 1956; Levine, 1966, 1970; Markant & Gureckis, 2014a; Millward & Spoehr, 1973; Trabasso & Bower, 1964), rule learning (Taplin, 1975; Tweney et al., 1980; Wason, 1960; Wason & Johnson-Laird, 1972), card selection (Cosmides, 1989; Cox & Griggs, 1982; Evans, 1982; Evans & Lynch, 1973; Evans, Newstead, & Byrne, 1993; Jones & Sugden, 2001; Tweney & Doherty, 1983; Wason, 1968; Wason & Johnson-Laird, 1972), contingency judgment (Alloy & Tabachnik, 1984; Arkes & Harkness, 1983; Crocker, 1981; Nisbett & Ross, 1980; Schustack & Sternberg, 1981; Shaklee & Mims, 1981, 1982; Ward & Jenkins, 1965), word learning (Siskind, 1996; Xu & Tenenbaum, 2007), and personality perception (Hodgins & Zuckerman, 1993; Schwartz, 1982; Snyder & Swann, 1978a, 1978b; Snyder & Campbell, 1980; Snyder, 1981; Strohmer & Newman,

1983; Trope & Bassok, 1982, 1983; Trope, Bassok, & Alon, 1984).

What do people do instead of strictly falsify? The typical approach is to construct *positive tests*, which are queries that are most likely to produce affirmative or confirmatory evidence if the learner's current hypothesis is true (Mynatt, Doherty, & Tweney, 1977; Nickerson, 1998; Shaklee & Fischhoff, 1982; Wason, 1960). Strictly speaking, a positive test strategy is not necessarily inconsistent with the *logic* of falsification: if the feedback is negative when the learner's hypothesis predicts it to be positive, that hypothesis will be falsified. The frequency with which a positive test also represents a falsificationist test depends on the relationship between the current hypothesis and the true hypothesis (Klayman & Ha, 1987, 1989), but except in special cases the two are not equivalent. Regardless, people are not motivated by falsification when seeking information: they tend to prefer positive tests, independent of whether they falsify.

A natural question to ask is why human learners deviate from falsificationism. One possibility is the fact that, in its simplest form, the logic of falsification focuses on a *single* hypothesis that must either be accepted or falsified. This kind of hypothesis testing shows up in the way that Fisher operationalised null hypothesis testing (Fisher, 1934; Lehmann, 2011). However, it is not obvious that attempting to falsify a single candidate hypothesis is the most useful way to learn about the world (Fedorov, 1972; MacKay, 1992). Another classic method is to consider two candidate hypotheses and conduct a test designed to discriminate between them. Thinking about hypothesis testing as a competition between two rival hypotheses underpins the Neyman version of hypothesis testing (Lehmann, 2011; Neyman & Pearson, 1933).[1] Of course, there is nothing special about the number two: the same logic can be extended to arbitrarily many hypotheses. Indeed, considering many hypotheses simultaneously is quite natural in Bayesian and information theoretic approaches to hypothesis testing (Austerweil & Griffiths, 2008; Klayman, 1987; Navarro & Perfors, 2011; Oaksford & Chater, 1994). When formulated in those terms, the goal of hypothesis testing shifts away from learning about any one hypothesis, and more towards selecting queries that will yield the most information about the identity of the true hypothesis.

In the psychological literature, Bayesian and information theoretic approaches to intuitive hypothesis testing are now commonplace (Austerweil & Griffiths, 2008; Klayman, 1987; Navarro

& Perfors, 2011; Oaksford & Chater, 1994). When viewed from the perspective of these approaches, there is considerable justification for learners to seek positive evidence because an affirmative response from the world/teacher is usually more informative than receiving a negative result (Austerweil & Griffiths, 2008; Navarro & Perfors, 2011; Oaksford & Chater, 1994). Different authors use slightly different approaches to arrive at the result, but in all cases this conclusion relies on the critical assumption that most of the natural hypotheses in the world return affirmative results to only a minority of possible queries. An example of such a hypothesis about natural numbers would be the hypothesis MULTIPLES OF 10: most numbers are not multiples of ten so the expected response to a query about a random natural number would be No. This assumption is sometimes referred to as a "rarity of yes" or "sparsity of hypotheses" assumption. For the purposes of this paper, we follow Tenenbaum and Griffiths (2001) and refer to it as the *size* of the hypothesis, which we define as the proportion of possible queries for which the hypothesis predicts an affirmative response. Thus, the hypothesis MULTIPLES OF 10 has a size of 10%. This concept of hypothesis size is orthogonal to the number of possible candidate hypotheses – in this example the set of possible rules for grouping integers is unbounded but each hypothesis has a specific size based on the proportion of affirmative responses to possible queries.

All of the information-theoretic approaches to hypothesis testing rely on the assumption that the average size of hypotheses is small. At one extreme, Austerweil and Griffiths (2008) discuss the situation when all hypotheses are as small as possible, being consistent with only one possible query or test. In the most general case that we are aware of, Navarro and Perfors (2011) find that only when the average size of hypotheses is less than 50% is it informationally optimal to perform a positive test. In fact, the information theoretic justification for seeking positive evidence not only vanishes when the small-hypotheses assumption is removed, but reverses if hypotheses are assumed to be large. That is, if hypotheses on average are large rather than small, seeking negative evidence is a better strategy.

The intuition behind this is readily understandable in visual terms. The left panel of Figure 1 is a schematic depiction of a situation in which most hypotheses are small. In such a case, a random piece of positive evidence will be inconsistent with and eliminate most of hypotheses, making it

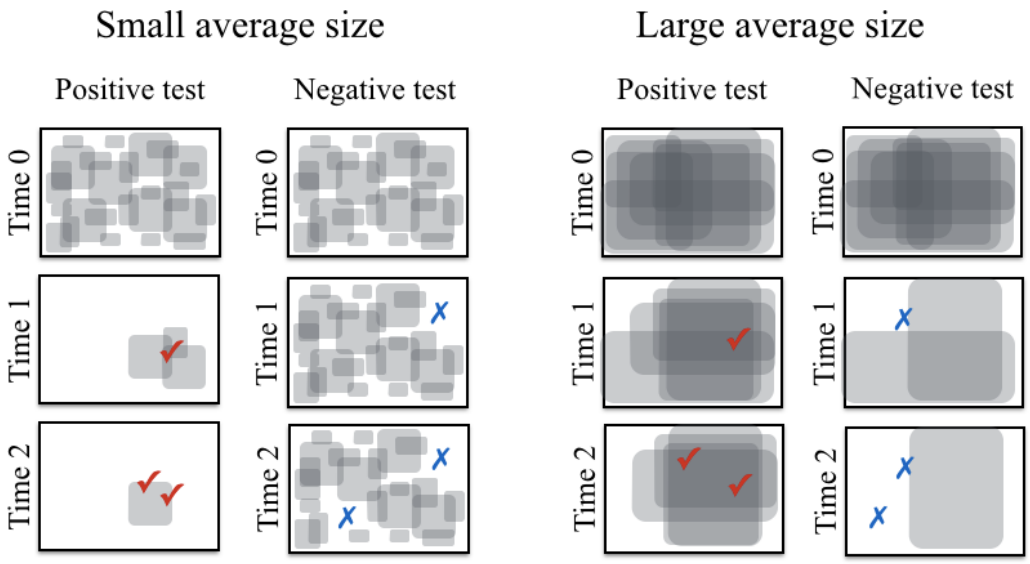


Figure 1. Schematic view illustrating why positive tests are better when the average size of hypotheses is small and negative tests are better for large hypotheses. Each of the light grey boxes represents a hypothesis in a space, and moving from top to bottom shows successive positive (check) or negative (x) queries. On the left panel is a space where the average size is small. When that is the case, each positive test eliminates most of the hypotheses so it is possible to identify the correct one with fewer queries when using positive tests and not negative tests. On the right panel, the opposite is true: the average size of the hypotheses is large, so each negative test eliminates most hypotheses and leads to identifying the correct one more quickly than positive tests.

possible to identify the correct hypothesis within a small number of tests. By contrast, negative evidence will eliminate few hypotheses, since most hypotheses are consistent with most negative evidence. When most hypotheses are large, shown on the right panel of Figure 1, the effect flips. Now most hypotheses give a Yes answer to most tests, so *negative* evidence is the type that leads to identifying the correct hypothesis more quickly.

Another way to understand why the utility of positive tests depends on the size of hypotheses in the space is to consider it quantitatively, casting the problem in Bayesian terms. The posterior probability $P(h|r,q)$ that hypothesis h is true given that query q was posed and response r received can be written as

$$P(h \mid r,q) = \frac{P(r \mid q,h)P(h)}{P(r \mid q)}$$

where $P(h)$ describes the learner's prior degree of belief in the hypothesis (assuming $P(h) = P(h|q)$) and the likelihood $P(r|q,h)$ equals 1 if hypothesis h predicts the response r and 0 if it does not.

The critical term is the denominator, $P(r|q)$, which describes the learner's overall belief about the probability of receiving response r, averaged across all possible hypotheses that the learner might entertain. If $P(r|q)$ is low then the overall boost to the learner's belief in hypothesis h is maximized. This occurs when the outcome r to query q is rare among all possible hypotheses, because that outcome produces the strongest evidence for hypotheses consistent with it and also eliminates the many hypotheses that were inconsistent, as pictured in Figure 1. As a consequence, if most hypotheses are small, then Yes is the rare outcome and selecting queries that are optimized for seeking confirmatory responses is the strategy that maximizes the evidentiary value of the typical response. Of course, if hypotheses are large, this logic is reversed and it becomes optimal to employ negative tests rather than positive ones. The higher utility of rare evidence occurs not only in these analyses of hypothesis testing but also in many inference problems (McKenzie & Chase, 2012; McKenzie & Mikkelsen, 2000).

The information theoretic analyses that advocate for confirmatory evidence rely on the assumption that most hypotheses are small. Is this assumption sensible? Oaksford and Chater (1994) justify it empirically, pointing out that most natural categories tend to be quite small relative to the size of the domain to which they belong: for instance, only a small proportion of animals are dogs. Investigations of the structure of natural categories bear this out (De Deyne et al., 2008), and to the extent that hypotheses are constructed from a consideration of natural categories people will therefore tend to consider small hypotheses as much as possible. Navarro and Perfors (2011) justify this assumption theoretically, showing that it is inherently easier to form coherent "family resemblance" categories with small sets of entities, and therefore a learner who seeks to form coherent categories should end up with small hypotheses.

As this discussion illustrates, there are good reasons to think that most hypotheses in the real world will probably tend to be small. However, the fact that this pattern should hold in general does not preclude the possibility that particular learning problems should violate them. For instance, suppose you were trying to learn the category of MOTILE ANIMALS among other similarly sized possible categories. Given that the majority of animals are motile, the most informative evidence would come from identifying examples of animals that cannot move (e.g., coral).

Do people behave as predicted by the information-theoretic analyses by switching to request negative evidence when the hypotheses are large? This paper addresses this question by departing from previous work and manipulating the size of hypotheses across conditions. If people tend to seek positive evidence because such evidence is more informative, we should expect this preference to disappear or indeed reverse when the learner knows that the hypotheses are large (as in the MOTILE ANIMALS example). If the preference for positive evidence is simply an irrational bias, then there is no reason to expect any such shift when the size of hypotheses changes. In order to avoid the difficult inferential step of determining if each test was confirmatory or not, we designed a task in which people were allowed to specifically request positive or negative evidence directly. Our experimental evidence suggests that people are quite sensitive to hypothesis size when gathering information to test hypotheses.

Method

The purpose of this experiment was to present participants with a realistic situation in which they could ask for positive evidence or negative evidence. We also wanted to be able to vary the size of the hypotheses in a natural way. This was accomplished using a task loosely based on the "Battleships" game (Gureckis & Markant, 2009; Markant & Gureckis, 2012, 2014b) in which participants were given a grid with five unique ships. Their job was to guess the location of the ships based on asking for evidence in the form of positive evidence (HITS) and negative evidence (MISSES). People in different conditions were shown ships of different sizes, which naturally changed the size of the hypotheses they were considering.

Participants

The initial data collection consisted of posting an online ad on Amazon's Mechanical Turk for 500 participants to complete the experiment; complete data was collected from 481 participants. Due to a coding error, the ship sizes from five of the nine conditions were calculated incorrectly and did not match the intended hypothesis sizes. The data from those conditions was discarded (255 participants) and a second online ad was posted for 250 new participants who were randomly assigned to one of the discarded five conditions. Complete data was collected for 180 additional

participants; 16 participants were eliminated who participated twice. Of the 390 total participants 9 were excluded because they failed to make at least one evidence request in each game. An additional 15 participants were excluded for making more responses in either game than 2.5 standard deviations above the mean (mean: 35, sd: 34). Of the remaining 366 participants included in all analyses, 35% were female. They ranged in age from 18 to 67 with a mean of 31. 69% of the participants were from the USA, 22% were from India, and the remainder were from 14 other countries. The mean completion time was 15 minutes, and participants were compensated US$0.50.

Materials & Procedure

The battleships game took place on a 20 by 20 grid partially covered by 5 grey rectangles ("ships") of different sizes with fixed orientations. No ships within a condition were identical (Figure 12 in the appendix shows the exact ships used in each condition). Figure 2 shows a sample board with an arrangement of ships. Participants were able to click and drag the ships to any location on the grid, so long as each ship remained within the grid and no ships overlapped with each other. They were told that their goal was to position the grey ships in their correct positions. The correct positions were randomly selected from a large set of plausible configurations identified by participants in a pilot experiment.[2]

In order to discover the correct configuration, participants could click on one of two buttons labelled Generate hit or Generate miss. If they clicked on the HIT button, the computer would randomly select (without replacement) one of the 400 squares within the grid that was covered by the correct configuration of ships and marked it with a red square. Conversely, if they clicked on the MISS button, the computer revealed a random location that was not contained within the correct configuration of ships and marked it with a blue square. The position of each ship was recorded after each evidence request but only analyzed to compute performance measures and the response time between each request was recorded but not analyzed. No other measures were collected.

During the task participants were instructed to move the grey rectangles to the positions that indicated their best guess of the correct locations of the ships. In order to encourage participants to use the rectangles in this fashion, if the current configuration was inconsistent with any of the

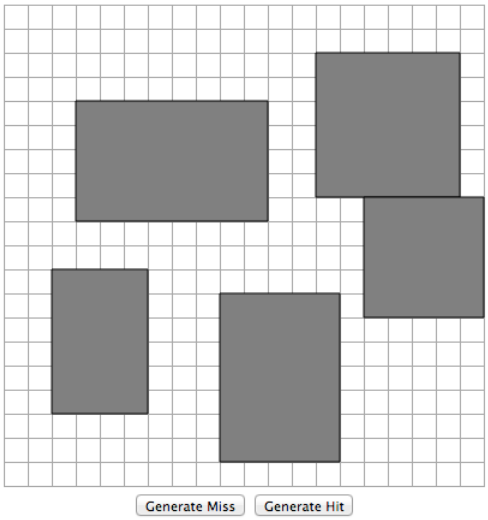
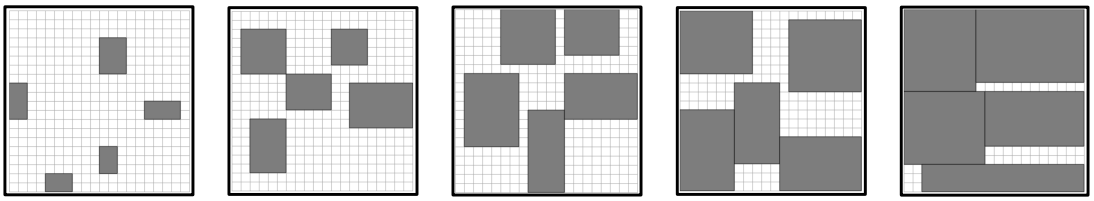


Figure 2. The experiment as it appeared to participants before making requests. The grey rectangles representing the ships are initially randomly positioned and participants can click on them to drag them to where they think the ships are truly located. They can also ask for as many MISSES (negative queries) or HITS (positive queries) as they want by clicking on the buttons at the bottom: asking for a miss reveals in a blue grid cell where none of the ships are, and asking for a hit reveals a red grid cell where one of the ships is.

revealed HITS or MISSES they were prevented from continuing until the rectangles were consistent with all the available evidence. People could ask for as many HITS and/or MISSES as they wanted until they were satisfied. At that point they would click Done to complete the task. After clicking Done, people were shown the position of the hidden ships along with their final guesses. They were also given a score based on the sum of squared distance between the top-left corner of their final guesses and the correct positions, divided by the number of information requests.

All participants began by playing a practice version of the game in which they were instructed to familiarize themselves with the setup, including pressing HIT and MISS at least once, and practice moving the ships around. After that, they played two games each, both within the same experimental condition. The practice version of the game was not analyzed.

Participants were randomly assigned to one of nine experimental conditions in which the shape of the ships was manipulated such that the portion of the grid covered by the ships ranged from 10% to 90% in steps of 10%. A subset of these conditions are visualized as game boards in Figure 3, and the complete list of ship sizes for each condition are included in the appendix. The



size = 10% size = 30% size = 50% size = 70% size = 90%

Figure 3. An illustration of five of the nine hypothesis size conditions. The configuration of ships in the first frame covers 10% of the grid and each subsequent frame covers 20% more than the next.

area of the grid covered by ships is directly proportional to the size of the hypotheses in a condition because the area determines the ratio of positive and negative evidence. Therefore the size of each hypothesis within a condition was the same and the size of hypotheses ranged from 10% to 90% positive across the nine conditions.

The total number of possible hypotheses was 5^{400} in each condition but the set of legal candidate hypotheses in which no ships overlapped was not fixed across conditions. In conditions with smaller ships there were many more legal candidate hypotheses than in conditions with larger ships, since there were many more situations in which no ships overlapped.

Results

Overall, performance in the task is quite good: after an average of 29.9 requests for HITS or MISSES per game, the final solutions have an average overlap of 87% with the correct configurations.[3] These analyses focus on the evidence requests people made across conditions, but a detailed breakdown of the average number of requests and the initial and final overlap across conditions, as well as any differences between the first and second board, is included in the appendix. In order to avoid over-weighting responses from participants who made many requests we calculate a single "positive preference" score for each person. This is defined as the percent of their requests that are for positive evidence (HITS) across both games.

To what extent does hypothesis size shape the preference for positive evidence? As Figure 4 demonstrates, there is a clear linear relationship between hypothesis size and the degree to which people prefer positive evidence. In the 10% condition the average preference for positive evidence is 86%, whereas in the 90% condition it is only 36%. Very clearly, the smaller the hypotheses, the

stronger the preference for positive evidence. To quantify this effect we analyze the data using a Bayesian linear regression implemented in JAGS (Plummer, 2003). The solid line in Figure 4 reflects the regression line estimated by the model, and the dotted lines show 95% posterior predictive intervals. The estimated slope of the regression is -0.63 (95% credible interval: $[-0.71, -0.55]$) indicating that each 10% increase of the hypothesis size reduces the average percent of positive requests by 6.3%.

Exploratory analysis of individual differences

This pattern of aggregate behavior suggests that people are highly sensitive to hypothesis size, but it may not necessarily reflect the patterns of each individual (Estes, 1956). Is each person's preference for positive information proportional to the hypothesis size? If so, we expect that all

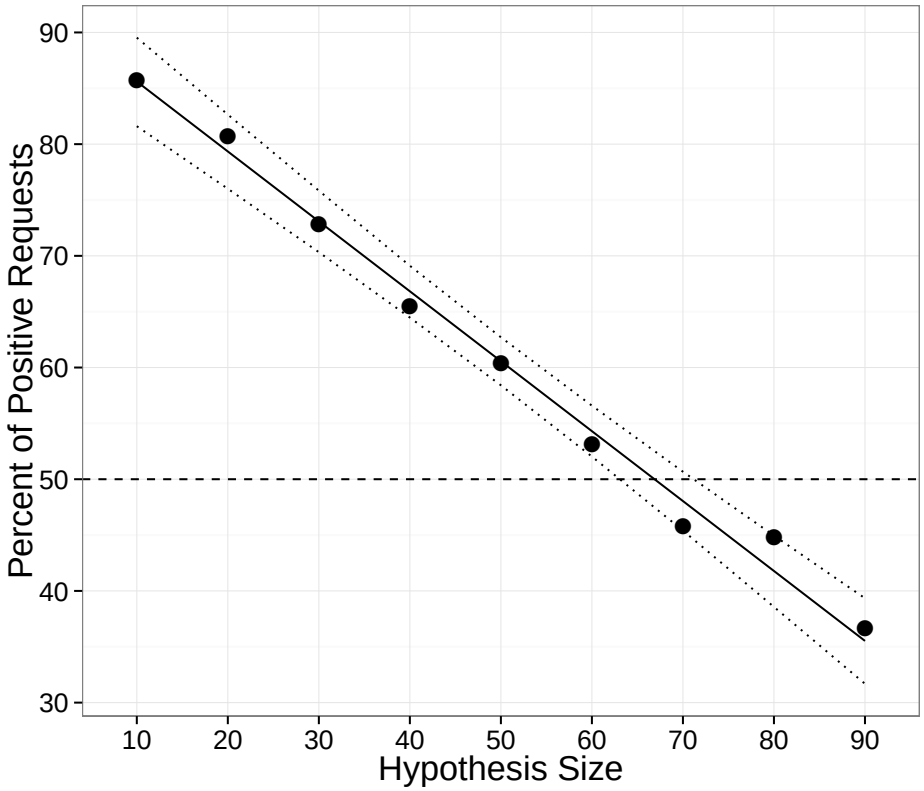


Figure 4. The percent of positive evidence requests (y axis) decreases as hypotheses become larger (x axis). Black dots show the average across participants within each condition. The solid black line shows the regression line, along with Bayesian 95% posterior predictive intervals for the mean for each condition.

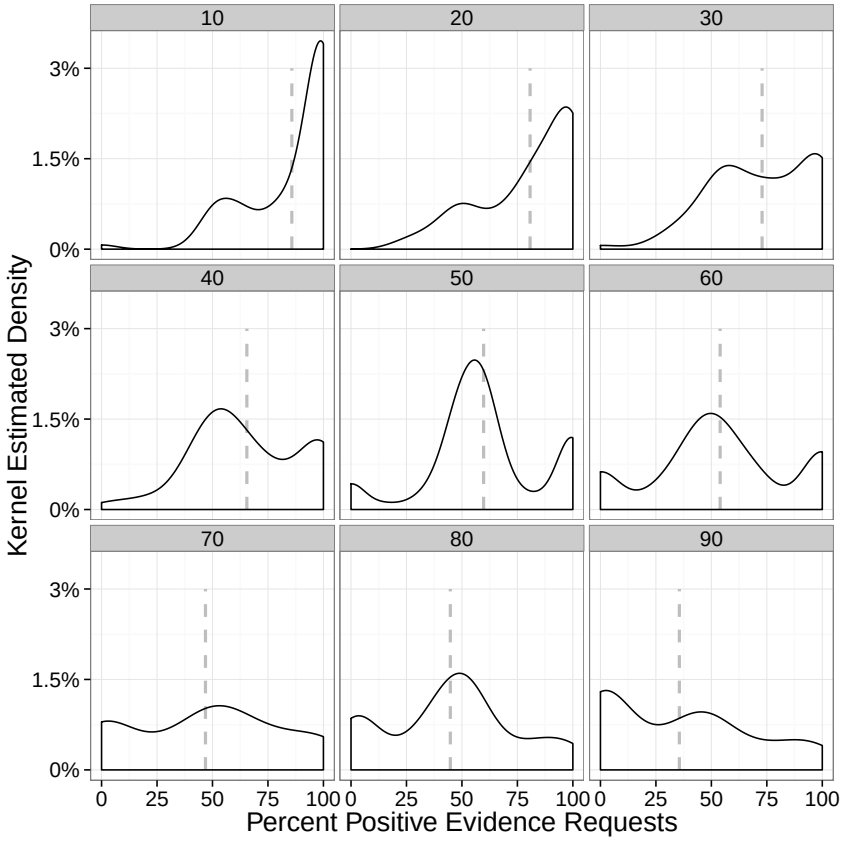


Figure 5. The distribution of the percent of positive evidence requests per game (x axis) is not unimodal nor centered at the mean (dashed grey line) of each condition. The multi-modal pattern in all conditions suggests people are using a mixture of strategies that produce very different percent of positive evidence requests. The distributions are smoothed using a kernel density estimator.

or most of the individuals in a given hypothesis size condition show similar behavior, with positive preference scores distributed with a single peak centered at the mean probability for the condition. Alternatively, it may be that there are only a few qualitatively distinct patterns of behavior (e.g., always ask for a HIT or always ask for a MISS) and different conditions have different mixtures of people following each pattern. These two possibilities would both produce the results in Figure 4 but would reflect very different underlying processes.

We investigate this issue by computing the percent of positive evidence requests within each game for each participant. The distribution of proportion of positive evidence requests, shown in Figure 5, are neither unimodal nor centered at the mean probability of each condition (indicated with a dashed grey line). In fact, the distribution in each condition is multimodal. Most people

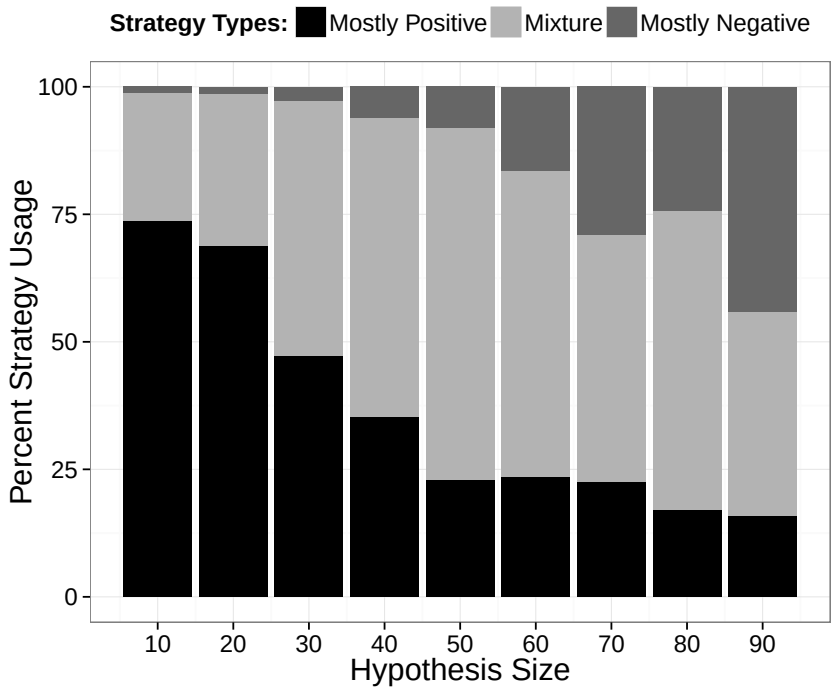


Figure 6. Proportion of people following each strategy (y axis) as a function of hypothesis size (x axis). More people use a `Mostly positive` strategy (requesting HITS more than 75% of the time, shown as a black bar) when hypotheses were small. For larger hypotheses, more people use a `Mostly negative` strategy (requesting HITS less than 25% of the time, shown as a dark grey bar) or a `Mixture` strategy (a light grey bar).

appear to either mostly ask for HIT, mostly ask for MISS, or evenly mix the two. This suggests that the overall sensitivity to hypothesis size in Figure 4 results from different mixtures of strategies rather than everyone slightly changing a single strategy.

To test whether the three different strategies suggested by Figure 5 actually do differ between conditions, we assign each game to one of the strategies. Those with more than 75% of positive requests are classified as `Mostly positive`, those with less than 25% are `Mostly negative`, and the rest are classified as using a `Mixture` strategy.[4] Overall, 264 games are classified as `Mostly positive`, 115 as `Mostly negative`, and 363 as a `Mixture`. Figure 6 displays the proportion of each strategy across conditions. The results parallel the aggregate trends seen in Figure 4 and reveal that most people use a `Mostly positive` strategy when the hypotheses are small but do not use it when the hypotheses are large. The opposite pattern occurs for the `Mostly negative` strategy, it is only used when hypotheses are large. Interestingly, although the general trend in both Figure 4

and 6 is of a sensitivity to hypothesis size, it is also true that behavior for the largest hypotheses is not a perfect mirror image of the smallest ones: there is more of a preference for positive evidence than one would expect if it were. A chi-squared test on the overall proportion of `Mostly positive` and `Mostly negative` games suggests a significant bias toward the `Mostly positive` strategy ($\chi^2(1) = 58.6, p < 0.0005$). We address this residual positive bias in the discussion section.

The information theoretic analyses of hypothesis size and the utility of evidence types (Navarro & Perfors, 2011) suggest that among the conditions with small hypotheses, the `Mostly positive` strategy should be more effective (i.e., it should eliminate more candidate hypotheses than the `Mostly negative` strategy). These analyses also predict that this advantage should be reversed in conditions with large hypotheses. If this prediction holds, we expect that on average people who follow the `Mostly positive` strategy should do better than the `Mostly negative` strategy if they are in a condition with small hypotheses, but should do worse if they are in a condition with large hypotheses. Figure 7 shows the performance, as measured by the overlap of the final guesses with the correct locations, as a function of the number of guesses they made and the strategy used. The pattern of performance matches the predictions: people using the `Mostly positive` strategy perform better than people using the `Mostly negative strategy` among the conditions with small hypotheses but worse among conditions with large hypotheses. No clear performance advantage between strategies occurred in conditions with moderately sized hypotheses.

The most popular strategy overall and especially in the moderately sized hypotheses was the `Mixture` strategy but it is unclear what process is producing these near equal mixtures of positive and negative evidence requests. Are they the result of a simple heuristic such as alternating between positive and negative evidence requests? One method to characterize what process is driving the `Mixture` strategy is to compare specific statistics of the sequence of positive and negative evidence requests to random sequences of requests. The number of observed alternations on each `Mixture` strategy board was compared to the distribution of alterations produced by 1,000 simulated games with a matched number of requests and probability of requesting positive evidence. The alternation strategy, or similar heuristics, will produce many more alternations than expected by random chance. Using the arbitrary threshold of being more extreme than one standard deviation above

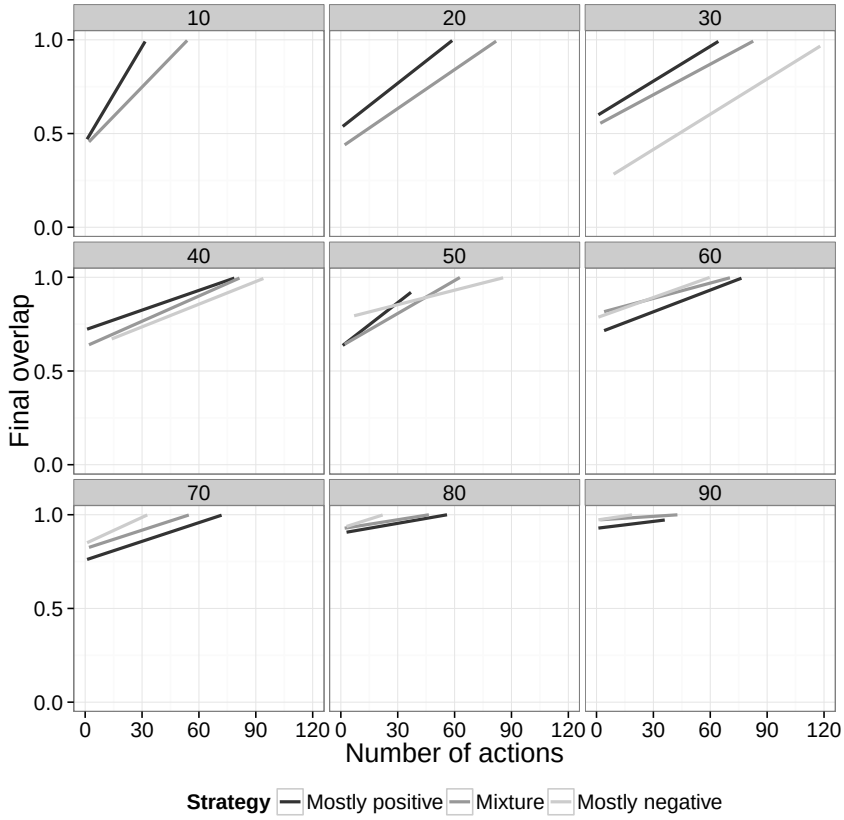


Figure 7. The best fitting linear relationship between the overlap of the final guess and the hidden locations of each ship (y axis) and the number of evidence requests (x axis) split by which strategy was used. People using the `Mostly positive` strategy had better performance in conditions with small hypotheses (top row) than other strategies but using the `Mostly negative` strategy resulted in better performance in conditions with large hypotheses (bottom row). The `Mostly negative` strategy is omitted from Conditions 10 and 20 because less than 2 people used that strategy in those conditions.

the average of simulated boards, 25% of the `Mixture` strategy boards had more alterations than the random process, suggesting those people were using something similar to an alternation strategy. However, using the same threshold, 60% of the actual `Mixture` strategy game boards had fewer alterations than expected by chance, suggesting people were making consecutive runs of the same evidence type and alternating rarely. Only 15% of the `Mixture` strategy boards had alternation counts consistent with random behavior.

Discussion

Several decades' worth of rational analyses have theorized that positive tests are sensible when a reasoner is confronted with hypotheses that are small relative to the size of the domain (Austerweil & Griffiths, 2008; Klayman & Ha, 1987, 1989; Navarro & Perfors, 2011; Oaksford & Chater, 1994). Since we live in a world where most hypotheses are small (De Deyne et al., 2008; Navarro & Perfors, 2011; Oaksford & Chater, 1994), these analyses imply that the preference for positive tests might be an appropriate adaptation rather than an irrational bias. Another implication also follows from these information theoretic approaches: if the preference for positive tests results from having small hypotheses, then one would expect this preference to be flipped when the hypotheses are large. Our work is the first to test this prediction (though see Hoffmann and Crott (2004) for a related manipulation). As predicted, we find that people's preference for positive evidence is highly sensitive to hypothesis size: positive tests are more frequently chosen when hypotheses are small; but when they are large, negative tests are selected more often.

This finding holds at both the aggregate and individual levels, but in ways that are revealing about the underlying processes involved. At an aggregate level, there was a clear linear relationship between the proportion of positive evidence requests and the size of the hypotheses being tested. But the aggregate pattern was not mirrored at the level of individual behavior because people did not ask for positive evidence at a frequency proportional to the size of the hypotheses. Rather, participants seemed to follow one of three strategies: asking for mostly positive, asking for mostly negative, or selecting the two evidence types approximately evenly. As hypotheses got larger, the proportion of people who asked for mostly positive evidence got smaller and the proportion asking for mostly negative evidence increased.

These results hide an interesting subtlety. Because each hypothesis in our experiments is defined in terms of five distinct ships that participants could move around on the screen, the "look and feel" of the task made it more natural to frame the hypothesis in terms of the cells occupied by the ships and not in terms of the "white space" cells left unoccupied by the ships. Indeed, when playing the game ourselves it felt quite unnatural to think of a configuration of ships in terms of the "white space" left unoccupied, even in the 90% condition. That said, it is mathematically possible

to reframe the hypotheses in terms of the space *not* occupied, thereby shifting the definition of positive and negative evidence. If that happened, the "70%" condition would in fact consist of a set of *small* hypotheses about the 30% of the grid occupied by white space. In practice, however, we found it almost impossible to think that way when doing the task, and given that the results from the actual 70% condition are not the same as the 30% condition, we suspect that participants found it equally unnatural. Indeed, we tried and failed to develop a different cover story (e.g., involving holes in a blanket) that naturally led people to think of the hypotheses in terms of the "white space"; we never succeeded because no matter what we did, it was always most natural to define the hypotheses in terms of the "figure" and not the ground. The structure of the task itself thus imposes a very strong framing effect in terms of how the evidence types and hypothesis sizes are defined.

One interesting aspect of our results is that behavior in the large hypothesis conditions is not a perfect mirror image of behavior in the the small hypothesis conditions. As seen in the aggregate behavior of Figure 4, people chose positive evidence 86% of the time in the smallest (10%) hypothesis condition, but in the largest (90%) condition they preferred negative evidence only 64% of the time. This suggests that although people are sensitive to hypothesis size, there is still a residual preference for positive tests – a preference that is most apparent when the hypothesis size is very large and such tests are most inappropriate. A similar effect is visible in the individual results in Figure 6: nearly 75% of people in the smallest condition follow the `Mostly positive` strategy, but just under 50% of people in the largest follow the `Mostly negative` one.

This residual preference for positive evidence – over and above the larger effect of a sensitivity to hypothesis size – is especially intriguing in light of the long-standing view that people have an irrational bias to select positive tests (Wason, 1960, 1968; Nickerson, 1998). Our results suggest there might be *some* bias, albeit a partially-defeasible one. This bias might arise from a sensitivity to the distribution of hypotheses in the world, in which most hypotheses are small (and hence favor positive tests). At a minimum, it is at least *affected* by the distribution of hypotheses: in three of the four conditions with large hypotheses the `Mostly negative` strategy was more popular than the `Mostly positive` strategy. The fact that this bias is defeasible suggests that the positive test

strategy may be less a product of an irrational bias and more an application of a general strategy that works well most of the time (since, after all, most hypotheses are small) but can be modified when people realize hypotheses are large and thus negative evidence is more useful.

This possibility is also consistent with the fact that people are good at our task but not more traditional hypothesis testing tasks. If people do well in our task because it enables them to override their default assumption that positive tests are better, then they should "fail" more on tasks that make it harder to recognize that such an assumption is inappropriate. Our task is quite straightforward in that respect. The visual nature of the battleship grid, with the rectangles showing the size of the ships, makes the size of the hypotheses immediately clear and highly salient. Moreover, it is easy to generate new hypotheses (by moving ships around) and the possibility of asking for negative rather than positive evidence is also salient (by having a Generate miss button).

The ease of generating negative evidence and the highly salient nature of hypothesis size on our task contrast with traditional hypothesis testing tasks like the 2-4-6 task (Wason, 1960) and the card selection task (Wason, 1968). In those, people have to generate their own hypotheses and the structure of the domain and the possible hypotheses is more opaque. If these apparently ancillary factors matter, one would expect that when they are removed from our task – perhaps by making it more like an abstract concept-learning one – people's performance should worsen. Indeed, in other work we investigated performance in a more abstract task and found that although people still showed the same overall sensitivity to hypothesis size, the effect was smaller (Langsford, Hendrickson, Perfors, & Navarro, 2014). This is also consistent with work showing that people perform much better on the Wason card task when the cover story is less abstract and more familiar or concrete (Cheng & Holyoak, 1985; Cosmides, 1989; Wason & Shapiro, 1971).

One concern might be that allowing direct requests for positive and negative evidence is not an ecologically valid task – that in the real world, people rarely ask for positive or negative evidence *per se*. Directly investigating whether this is empirically true would require an additional study, but our sense is that this kind of thing is not actually very unusual. For instance, people frequently ask each other questions like "What musicians do you [don't you] like to listen to?" These questions are especially common to and from children (e.g., "What animals [do / don't] live in a swamp?") On

a more practical level, using a task that allows participants to directly request a type of evidence avoids the difficult but otherwise necessary step of inferring if a particular query was intended to produce positive or negative evidence. The battleship task allows us to directly measure the type of evidence people want to use rather than trying to infer based on their query.

The degree to which the real world is more accurately captured by the battleship task – where hypothesis size is clear and requests are straightforward to generate – or more traditional hypothesis testing tasks is an open question not answered by this work. However, this research does suggest that when people struggle to optimally test hypotheses the first assumption should not be that they are inherently bad at selecting useful tests, but that they may be struggling to correctly perceive the size of hypotheses or to generate useful possible information requests. This view dovetails nicely with the large active learning literature in hypotheses testing that suggests in many domains people select test items that approximate information gain maximization (Markant & Gureckis, 2012; Nelson, McKenzie, Cottrell, & Sejnowski, 2010; Nelson & Movellan, 2001) or uncertainty reduction (Markant & Gureckis, 2014b). Overall, our results add to the growing evidence that suggests that a preference for positive tests may, at least in part, be sensitive *to* and sensible *for* the world in which we live.

Footnotes

[1]Indeed, although most undergraduate statistics textbooks collapse Fisher's approach and Neyman's view into a single "orthodoxy" of hypothesis testing, they disagreed with each other quite strongly at the time (Lehmann, 2011) for precisely this reason. Fisher's view was that one should seek only to accept or reject a single null hypothesis, whereas Neyman stressed the importance of having a well defined alternative against which the null could be contrasted. Neither Fisher nor Neyman viewed these as equivalent.

[2]The exact instructions were: "We are interested in making a game that has a variety of starting boards that consist of rectangles spread randomly inside a space. Your goal is to generate some of these boards. In this task you will be asked to position five rectangles within a grid so that they do not overlap. All five rectangles will initially be positioned in the same place but you should move them all to create an arrangement that looks and feels sort of random." We determined configurations this way rather than just randomly generating them for two reasons. The first is that when ships, and thus hypothesis sizes, were large, determining ship positions that do not overlap is a difficult computational problem. The second was that we wanted to get a typical range of plausible-feeling ship configurations, which was straightforward to do by having other participants generate them. This resulted in a set of 1721 different ship configurations (an average of 191 per condition) to randomly select from.

[3] Overlap is highly correlated with the sum of squared distance between guesses and locations ($r = 0.71$) but we use overlap because it was not strongly skewed by a few boards in which participants swapped the locations of ships.

[4]The 25/75 threshold is arbitrary but sensible given the task, and the qualitative finding is robust to a wide range of reasonable threshold values.

References

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, *91*(1), 112-148.

Arkes, H. R., & Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General*, *112*(1), 117-135.

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.

Bruner, J. S., Goodnow, J. J., & George, A. (1956). *A study of thinking.* John Wiley & Sons.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*(4), 391–416.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276.

Cox, J. R., & Griggs, R. A. (1982). The effects of experience on performance in Wason's selection task. *Memory & Cognition*, *10*(5), 496–502.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin*, *90*(2), 272-290.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134-140.

Evans, J. S. B. (1982). *The psychology of deductive reasoning.* Routledge & Kegan Paul.

Evans, J. S. B., & Lynch, J. (1973). Matching bias in the selection task. *British Journal of Psychology*, *64*(3), 391–397.

Evans, J. S. B., Newstead, S. E., & Byrne, R. M. (1993). *Human reasoning: The psychology of deduction.* Psychology Press.

Fedorov, V. V. (1972). *Theory of optimal experiments.* Elsevier.

Fisher, R. A. (1934). *Statistical methods for research workers.* Oliver and Boyd.

Gureckis, T. M., & Markant, D. B. (2009). Active learning strategies in a spatial concept learning game. In *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 3145–3150). Austin, TX: Cognitive Science Society.

Hodgins, H. S., & Zuckerman, M. (1993). Beyond selecting information: Biases in spontaneous questions and resultant conclusions. *Journal of Experimental Social Psychology*, *29*(5), 387–407.

Hoffmann, C., & Crott, H. (2004). Effects of amount of evidence and range of rule on the use of hypothesis and target tests by groups in rule-discovery tasks. *Thinking & reasoning*, *10*(4), 321–354.

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, *50*(1), 59–99.

Klayman, J. (1987). *An information theory analysis of the value of information in hypothesis testing* [Working paper no. 171]. University of Chicago, Graduate School of Business, Center for Decision Research.

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.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge University Press.

Langsford, S., Hendrickson, A. T., Perfors, A. F., & Navarro, D. J. (2014). People are sensitive to hypothesis sparsity during category discrimination. In *Proceedings of the 36th annual conference of the Cognitive Science Society* (pp. 2531–2536). Austin, TX: Cognitive Science Society.

Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics.* Springer.

Levine, M. (1966). Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology*, *71*(3), 331-338.

Levine, M. (1970). Human discrimination learning: the subset-sampling assumption. *Psychological Bulletin*, *74*(6), 397-404.

MacKay, D. J. (1992). *Bayesian methods for adaptive models* (Unpublished doctoral dissertation). California Institute of Technology.

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. (2014a). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94-122.

Markant, D. B., & Gureckis, T. M. (2014b). A preference for the unpredictable over the informative during self-directed learning. In *Proceedings of the 36th annual conference of the Cognitive Science Society* (pp. 958–963). Austin, TX: Cognitive Science Society.

McKenzie, C. R. M., & Chase, V. M. (2012). Why rare things are precious: The importance of rarity in lay inference. In P. M. Todd, G. Gigerenzer, & T. A. R. Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 309–334). Oxford: Oxford University Press.

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of hempel's paradox of confirmation. *Psychonomic Bulletin & Review*, *7*(2), 360–366.

Millward, R. B., & Spoehr, K. T. (1973). The direct measurement of hypothesis-sampling strategies. *Cognitive Psychology*, *4*(1), 1–38.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The Quarterly Journal of Experimental Psychology*, *29*(1), 85–95.

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

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

Nelson, J. D., & Movellan, J. R. (2001). Active inference in concept learning. *Advances in Neural Information Processing Systems*, 45–51.

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society*, *231*, 289–337.

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

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Prentice-Hall.

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

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353.

Plummer, M. (2003). Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (pp. 20–22).

Popper, K. (1934/1959). *The logic of scientific discovery.* Hutchinson.

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

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, *110*(1), 101-120.

Schwartz, B. (1982). Reinforcement-induced behavioral stereotypy: How not to teach people to discover rules. *Journal of Experimental Psychology: General*, *111*(1), 23-59.

Shaklee, H., & Fischhoff, B. (1982). Strategies of information search in causal analysis. *Memory & Cognition*, *10*(6), 520–530.

Shaklee, H., & Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Development*, *52*, 317–325.

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(3), 208-224.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1), 39–91.

Snyder, M. (1981). Seek, and ye shall find: Testing hypotheses about other people. In *Social cognition: The Ontario symposium* (Vol. 1, pp. 277–304).

Snyder, M., & Campbell, B. (1980). Testing hypotheses about other people the role of the hypothesis. *Personality and Social Psychology Bulletin*, *6*(3), 421–426.

Snyder, M., & Swann, W. B., Jr. (1978a). Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology*, *14*(2), 148–162.

Snyder, M., & Swann, W. B., Jr. (1978b). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, *36*(11), 1202-1212.

Strohmer, D. C., & Newman, L. J. (1983). Counselor hypothesis-testing strategies. *Journal of Counseling Psychology*, *30*(4), 557-565.

Taplin, J. E. (1975). Evaluation of hypotheses in concept identification. *Memory & Cognition*, *3*(1), 85–96.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(04), 629–640.

Trabasso, T., & Bower, G. (1964). Presolution reversal and dimensional shifts in concept identification. *Journal of Experimental Psychology*, *67*(4), 398-399.

Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology*, *43*(1), 22-34.

Trope, Y., & Bassok, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology*, *19*(6), 560–576.

Trope, Y., Bassok, M., & Alon, E. (1984). The questions lay interviewers ask. *Journal of Personality*, *52*(1), 90–106.

Tweney, R. D., & Doherty, M. E. (1983). Rationality and the psychology of inference. *Synthese*, *57*(2), 139–161.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., & Arkkelin, D. L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, *32*(1), 109–123.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *19*(3), 231-241.

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.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* Harvard University Press.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *The Quarterly Journal of Experimental Psychology*, *23*(1), 63–71.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245-272.

Appendix: Learning across boards

In this appendix we address the degree to which people learned or changed their strategy from the first board to the second. Overall, the proportion of positive evidence requests has a strong positive correlation ($r = 0.75$) between the first and second board. Figure 8 shows the distribution of positive evidence requests across each board for all conditions. The correlation between the first and second board within each condition ranged from a minimum of $r = 0.44$ in condition 30% to a maximum of $r = 0.77$ in condition 70%. This high correlation produces distributions of positive evidence requests that are very similar across the first and second board, as seen in Figure 9.

The similarity between evidence requests in the first and second board results in 75% of participants using the same request strategy on both boards. Figure 10 shows a heat map indicating the proportion of people that use each possible strategy combination across the two boards. The most probable cells in each condition fell along the major diagonal, which indicates that the participant used the same strategy on both boards. Though no strong pattern emerges, the trend appears to be a slight shift toward using the more useful strategy on the second board: a shift toward using the `Mostly positive` strategy in small hypothesis conditions and toward the `Mostly negative` strategy in the large hypothesis conditions.

However the slight shifts in strategy usage do not produce large shifts in the overall proportion of strategies used in each condition. Figure 11 shows the distribution of strategies people use split by board and the pattern suggest that there are not large overall differences from the first to the second board for any condition.

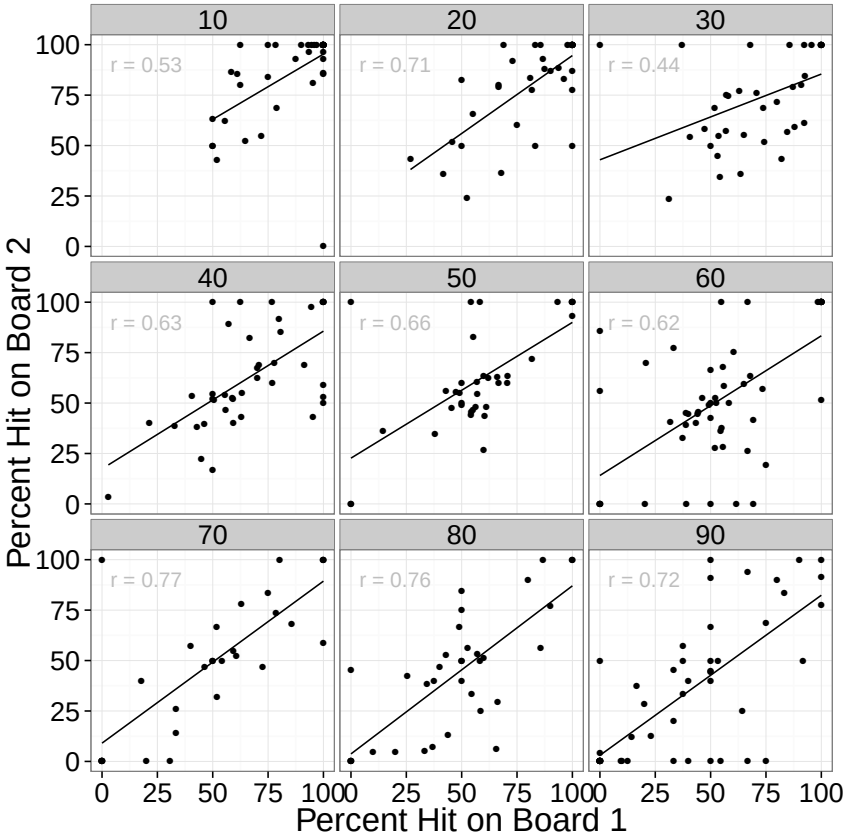


Figure 8. The proportion of positive evidence requests on the second board (y axis) are positively correlated with the proportion on the first board (x axis) across all hypothesis size conditions. Each point corresponds to the proportion of positive evidence requests for one participant on the first and second board. The black lines indicate the best fitting line for each condition. The correlation between first and second board was positive in all conditions and ranged from 0.44 to 0.77.

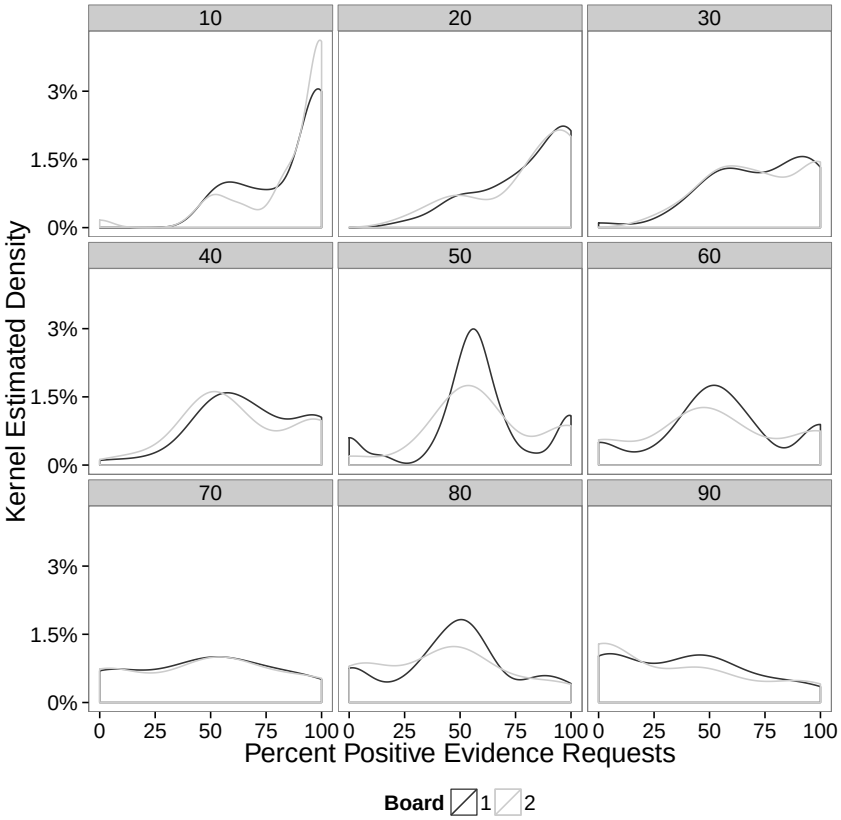


Figure 9. The distribution of the percent of positive evidence requests per game (x axis) is similar across the first and second board. The distributions are smoothed using a kernel density estimator.

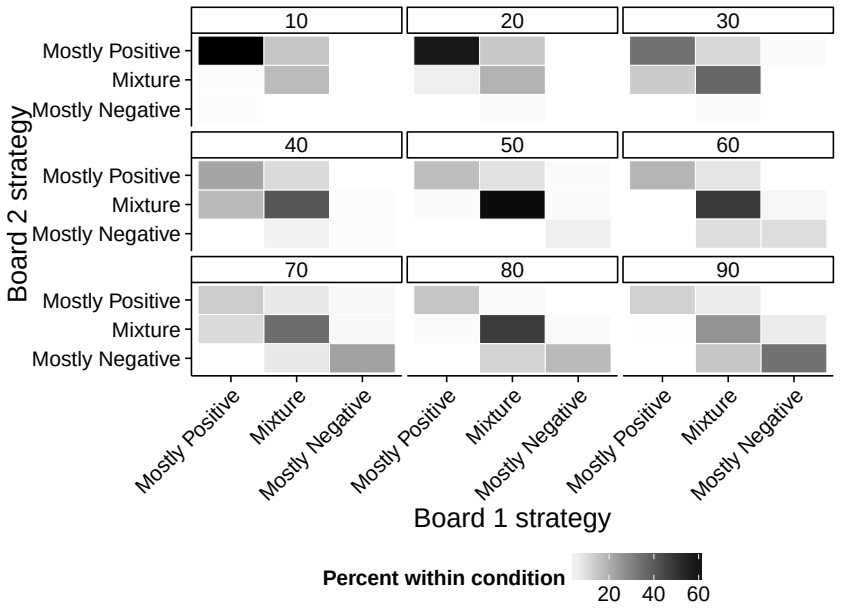


Figure 10. A heatmap showing the strategy each participant used on the first (x axis) and second board (y axis). Cells along the main diagonal indicate participants who used the same strategy on both boards, cells below the diagonal are the proportion of participants who shifted toward requesting more negative evidence requests, and cells above are participants who shifted toward requesting more positive evidence. 75% of participants across all conditions used the same strategy on both boards and participants who did shift strategies do not show a consistent pattern beyond a reluctance to shift directly between the `Mostly positive` and `Mostly negative` strategies.

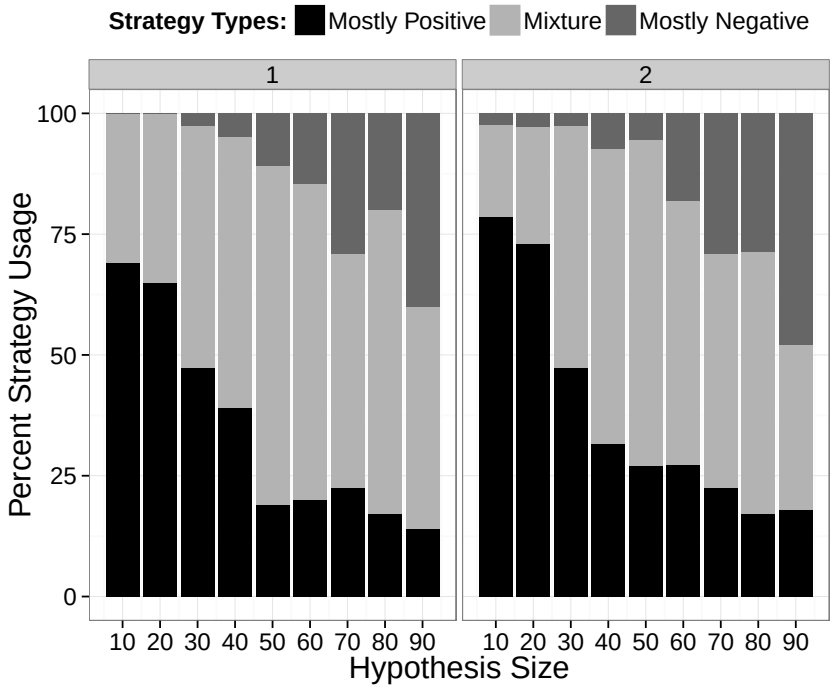


Figure 11. Proportion of people following each strategy (y axis) as a function of hypothesis size (x axis) split by board number. The distributions of strategy use are very similar across the two boards. Strategies were determined following the same procedure as in Figure 6.

Condition	Ship Sizes: Width by Height					Initial Overlap	Final Overlap	Actions
10	2x3	2x4	3x2	3x4	4x2	10.7 (9.4)	80.1 (24.4)	24.8 (15.6)
20	3x3	4x4	4x5	5x4	5x3	24.6 (10.0)	74.0 (25.8)	32.0 (27.2)
30	4x4	4x6	5x4	5x5	7x5	32.3 (9.9)	80.5 (21.5)	42.6 (30.3)
40	4x6	5x5	5x7	6x6	8x5	42.6 (9.7)	85.7 (15.8)	47.1 (30.0)
50	8x5	6x8	5x8	7x6	6x5	54.5 (9.1)	83.6 (14.8)	33.3 (22.5)
60	7x8	5x6	7x6	8x5	9x8	62.9 (8.8)	89.0 (11.7)	36.2 (26.5)
70	9x7	9x6	6x9	8x8	5x9	72.7 (7.5)	90.5 (10.4)	25.8 (21.8)
80	8x8	11x6	7x9	5x11	9x8	83.0 (5.6)	96.1 (5.1)	20.0 (15.5)
90	8x9	9x8	11x6	12x8	18x3	90.9 (2.5)	97.7 (2.9)	11.3 (11.2)

Figure 12. Ship sizes across all conditions. Initial and final overlap for each condition. Means across both boards are reported (standard deviations shown in parentheses).