Hayes, B. K., Banner, S., Forrester, S. and Navarro, D. J. (2019).
Selective sampling and inductive inference: Drawing inferences
based on observed and missing evidence. *Cognitive Psychology, 113*
https://doi.org/10.1016/j.cogpsych.2019.05.003

Selective sampling and inductive inference:

Drawing inferences based on observed and missing evidence

Brett K. Hayes [1], Stephanie Banner [1], Suzy Forrester [1], Danielle J. Navarro [1]

[1] University of New South Wales

Running Head: Selective sampling and inductive inference

Please address correspondence to:

Brett K. Hayes
School of Psychology
University of New South Wales
NSW, 2052   AUSTRALIA
Email: B.hayes@unsw.edu.au

Abstract

We propose and test a Bayesian model of property induction with evidence that has been selectively sampled leading to “censoring” or exclusion of potentially relevant data. A core model prediction is that identical evidence samples can lead to different patterns of inductive inference depending on the censoring mechanisms that cause some instances to be excluded. This prediction was confirmed in four experiments examining property induction following exposure to identical samples that were subject to different sampling frames. Each experiment found narrower generalization of a novel property when the sample instances were selected because they shared a common property (property sampling) than when they were selected because they belonged to the same category (category sampling). In line with model predictions, sampling frame effects were moderated by the addition of explicit negative evidence (Experiment 1), sample size (Experiment 2) and category base rates (Experiments 3-4). These data show that reasoners are sensitive to constraints on the sampling process when making property inferences; they consider both the observed evidence and the reasons why certain types of evidence has not been observed.

Keywords: Inductive reasoning; Property inference; Categorization; Bayesian models

1. Introduction

People routinely make inferences based on samples of evidence. A personnel manager who wants to know which employee attributes are most likely to lead to job success may review the records of current employees. Attributes correlated with successful performance in this sample may be used to select future employees. However, the evidentiary value of such a sample depends on how it was *selected* (Hogarth, Lejarraga, & Soyer, 2015). In the personnel example, clearly some individuals such as those who were previously unsuccessful in their job application are missing from the sample. In some cases, the reasons for this exclusion will be important for the kinds of inferences that one draws. If there were time constraints on the selection process that meant only a relatively small pool of local applicants could apply, or if there was a selection bias that favored male applicants, this limits the inferences that can be drawn.

Determining how people factor such selective sampling into their inferences is crucial because in everyday experience, such sampling is the norm – we typically observe only part of a relevant sample of evidence, with many observations excluded. The overarching goal of the current work therefore was to examine how people make inductive inferences based on samples of evidence where some portion of the relevant data has been *censored* or systematically excluded from observation.

In property induction tasks people are presented with a sample of instances that share some novel property and asked to infer how far this property can be generalized. Such reasoning is recognized as a central topic in cognitive science and has close links to domains like judgment and decision-making (Hayes & Heit, 2018). Traditionally, cognitive models of induction have focused on how the *content* of the observed sample determines how people generalize novel properties. For instance, knowing that a rhino is similar to a hippo makes people more willing to believe that properties that are true of rhinos will also be true of hippos, while no such charity is afforded when generalizing from rhinos to koalas.

Research in this area has revealed many important principles that guide inductive generalization (see Feeney, 2017 for a recent review). As well as similarity, factors such as the diversity, typicality and quantity of evidence, have all been shown to impact inductive inference. Formal mathematical models have been developed to account for each of these effects (e.g., Heit, 2007; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Sloman, 1993).

Despite the ubiquity of data censoring, this issue has generally been neglected in cognitive theories of reasoning (see Hahn & Oaksford, 2007; Hsu, Horng, Griffiths, & Chater, 2016 for notable exceptions). This contrasts with work in fields such as statistics, machine learning, bioinformatics, and economics (e.g., Jessen, 1978; Little & Rubin, 2014). Models in each of these fields have been developed to explain how to make inferences based on both observed data and data that has been systematically excluded from evidence samples. Taking some of the ideas from this literature as an inspiration, we develop and test a new computational theory of property inference with censored samples.

1.1 Data Censoring, Sampling and Bayesian Reasoning

A growing body of evidence suggests that people’s inductive inferences are guided by their understanding of the sampling process – the method by which observations were generated or selected (e.g., Navarro, Dry, & Lee, 2012; Shafto, Goodman, & Griffiths, 2014; Tenenbaum & Griffiths, 2001; Voorspoels, Navarro, Perfors, Ransom, & Storms, 2015; Xu & Tenenbaum, 2007). Much of the previous work in this area has examined sampling processes as a form of social cognition: people reason differently when they believe data were selected by a helpful teacher than when data are selected randomly, or when deception is involved. This work has produced some notable findings. Ransom, Perfors and Navarro (2016) showed that adding positive evidence (e.g., discovering that eagles as well as hawks share a novel property) can either increase or decrease property generalization depending on whether the learner assumes that the evidence was sampled randomly or supplied by a helpful teacher. Likewise, the well-known effect of evidence diversity (Kary, Newell, & Hayes, 2018; Osherson et al., 1990), whereby a property shared by dissimilar category members (e.g., dogs and whales) is more likely to be generalized to a superordinate category (e.g., mammals) than a property shared by similar instances (e.g., dogs and cats), depends on the assumption that the observed instances were the result of helpful sampling (Hayes, Navarro, Stephens, Ransom, & Dilevski, in press). These findings however, are typically limited to situations where the relevant sampling mechanism is social or communicative in nature.

The current work takes the study of sampling processes in a novel direction, investigating how inductive inferences can be shaped by mechanistic sampling constraints that lead to the selective sampling of some instances but not others. We begin by outlining our theoretical framework, and discuss connections with related work later in the paper.

1.2 A Bayesian framework for property inference with censored data

The central claim made by Bayesian models of inductive reasoning and inductive generalization is that human inferences can be described as a form of statistical inference. The learner approaches the problem with a prior distribution $P(h)$ defined over some class of possible hypotheses, and has some theory of the world that specifies the likelihood $P(d|h)$ of observing data d if hypothesis h were true. The effect of sampling assumptions is captured in this framework by specifying different likelihoods for different sampling conditions: a helpful teacher does not select data d in a purely random way, for instance, so one should expect that the likelihood $P(d|h)$ should be different when evidence is presented by a helpful teacher than when the same evidence arrives by a random sampling process (e.g., Navarro et al., 2012; Sanjana & Tenenbaum, 2003; Tenenbaum & Griffiths, 2001). In social reasoning contexts, the sampling model can be very complicated, insofar as it depends on the learner's theory of mind in respect of the teacher (Shafto et al., 2014, Voorspoels et al., 2015; Xu & Tenenbaum, 2007) and beliefs about the reliability of information provided by the teacher (Bovens & Hartmann, 2003; Oaksford & Hahn, 2013).

A key theoretical innovation in the current work is that we consider a class of Bayesian inductive reasoning models in which the data *d* are subject to a *deterministic censoring mechanism* defined by a simple survivor function *S(d)*. For data that are not subject to censoring, *S(d) = 1*, whereas data subject to censoring the survivor function is *S(d) = 0*. If *P(d|h)* is the likelihood that data *d* would have been generated in an environment without censoring, then the probability of the learner observing this data in an environment where censoring operates is simply *P(d|h) S(d)*. Applying Bayes' rule, a learner who encounters data *d* and is aware that censoring process with survivor function *S* applies, will compute the posterior probability of hypothesis *h* as follows:

$$P(h|d, S) \propto S(d) P(d|h) P(h) \qquad (1)$$

This is illustrated schematically in Figure 1.

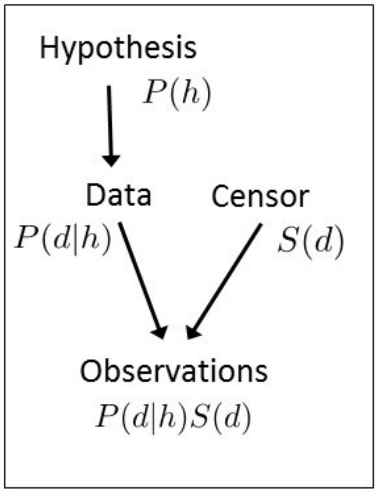


Figure 1. Outline of the Bayesian framework for inference with censored data

Though somewhat simplistic, this framework includes Bayesian strong sampling and weak sampling models (Tenenbaum & Griffiths, 2001) as special cases. Weak sampling occurs when the survivor function *S(d)*=1 for all *d*, and strong sampling occurs when *S(d)*=1 for all members of a category and 0 otherwise. Indeed, there is a sense in which the "censoring" framework reframes the core insight that the Tenenbaum and Griffiths (2001) model employed in a more general fashion. By doing so we are able to make a number of key predictions about how selective sampling of some instances and exclusion of others affects property inferences. Moreover, the framework is naturally extensible. While the current paper only considers scenarios where *S(d)* is either 0 or 1, it could be extended to cover *soft censoring* scenarios where *S(d)* can take on intermediate values, *conditional censoring* where the censoring mechanism depends on additional information about the state of the world *S(d|w)* (e.g., the detectability of stimuli might depend on the noisiness of an information channel that might change over time), or – with a slight change to the notation – *data transformation* scenarios in which *S* describes a mechanism in which some observations *d* are modified before being observed (e.g., extreme values are truncated to a narrow range).

1.2.1 The effect of sampling frames on property inference.

A central and novel prediction of the censored inference framework in Figure 1 is that the inferences that people make in a property induction task will depend on the survival function *S(d)*, even when the data *d* presented to people does not change. To this end, we devised an experimental program in which different groups of participants were exposed to identical data samples selected via sampling schemes that give rise to different survivor functions. Borrowing from the statistics literature (cf. Jessen, 1978), we use the term "sampling frames" to refer to this manipulation. A *category sampling* frame for example, refers to the case where instances are included in the sample because they belong to a particular category, with the members of other categories excluded. By contrast, a *property sampling* frame restricts inclusion in the evidence sample to instances that share some target property.

Consider a situation in which participants are shown a sample of instances that share a novel property (e.g., “small birds that have plaxium blood”) and asked to infer whether the property generalizes to other entities (e.g., large birds, mammals). Under a *category sampling* scheme, they would be told that the items were included in the sample because they are category members (small birds), implying that other animals were not eligible for inclusion, schematically illustrated in Figure 2a. In contrast, under a *property sampling* scheme exemplars were eligible for inclusion in the sample by virtue of the fact that they possess the property (plaxium blood), implying that plaxium negative animals were not eligible for inclusion, per Figure 2c.

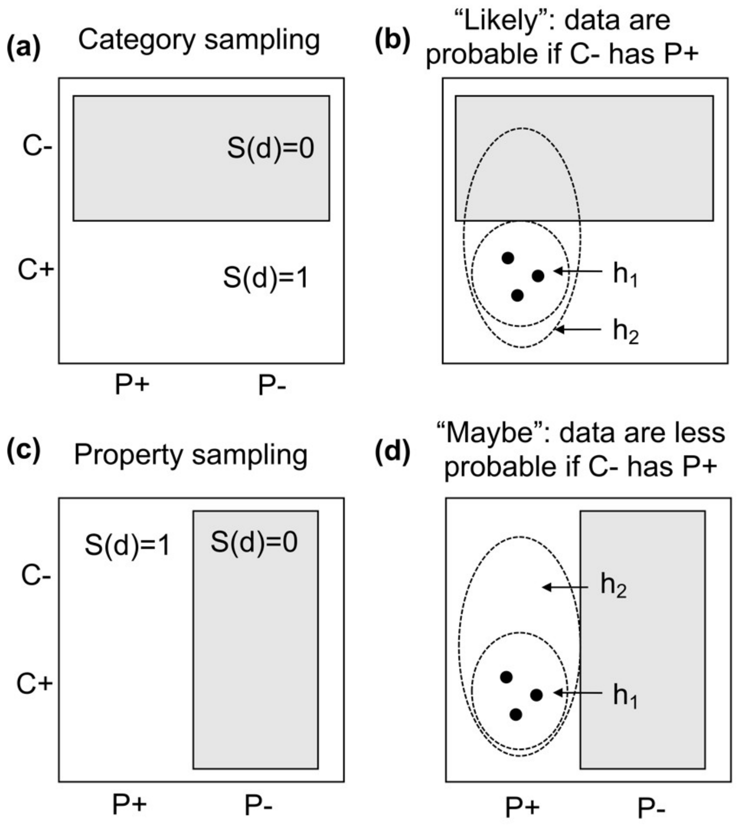


Figure 2. Schematic illustrations of category sampling (panel a) and property sampling (panel c). Possible observations are coded in terms of whether they belong to a target category (C+ or C-) and whether they possess the novel property (P+ or P-). Grey boxes denote the censor (S(d)=0) and unshaded areas are uncensored (S(d)=1). Panels b and d illustrate how the presence of the censor changes the relative plausibility of two hypotheses (h_1 and h_2, denoted with ellipses) about what observations are possible. See main text for detail.

The implication for inductive generalizations are sketched out in Figures 2b and 2d. Consider two possible hypotheses about the extension of the property P. According to hypothesis 1 (h_1), only the observed entities possess the property (e.g., only small birds possess plaxium blood) whereas hypothesis 2 (h_2) postulates that entities outside the category (e.g., large birds, mammals) can *also* possess plaxium blood. Under property sampling, a Bayesian reasoner will – typically, depending on the precise nature of the Bayesian model, consider h_1 more plausible than h_2. If there were animals besides small birds that could possess plaxium blood (as per h_2), we ought to have observed some in our sample. The fact that we do not see any such animals counts as implicit negative evidence, and accordingly is evidence for h_1 over h_2. In contrast, the same data collected under a category sampling scenario does not provide the same evidence: the absence of animals other than small birds can be explained by the operation of the censoring mechanism *S(d)*.

Although the illustrations in Figures 2b and 2d are schematic, they highlight the core Bayesian principle at work here: broadly speaking, the learner treats the data *d* as a random sample from the hypothesis *h*, operationalized by the ellipses in Figures 2b and 2d. The probability of observing any particular data set *d* is usually given by the likelihood $P(d \mid h)$ (e.g., selecting a location at random inside the ellipse), so “larger” hypotheses assign lower probability to the data (the size principle: Tenenbaum & Griffiths, 2001). This is what drives the learner’s preference for h_1 over h_2 in a property sampling scenario. However, when data censoring is relevant, the probability of observing particular data also depends on the censor. Observed data are sampled at random only from *uncensored* locations (where

S(*d*)=1). This makes no difference for property sampling, but it does matter for category sampling, because h_1 and h_2 now assign roughly the same probability to the observed data.

In short, a key prediction of this framework is that – all else being equal – people should be more willing to generalize a novel property in a category sampling scenario than in a property sampling situation. Later in the paper we will introduce a specific Bayesian model that instantiates these ideas in a precise, quantitative fashion, but for the empirical section we avoid doing so to highlight the fact that the key predictions emerge from the core framework and are not dependent on the particular instantiation we adopt later.

1.2.2 Experimental tests

There is some previous evidence suggesting that people treat category sampling differently to property sampling (Hayes, Banner, & Navarro, 2017; Lawson & Kalish, 2009). Lawson and Kalish (2009) for example, presented participants with a sample of animals (small birds) that shared a novel property and gave different groups cover stories that implied that the sample was selected either via category or property sampling. Those in the property sampling condition were less likely to generalize the property to other animals than those in the category sampling condition. Lawson and Kalish (2009) noted that these results could not be explained by existing models of induction (e.g., Osherson et al., 1990; Sloman, 1993), but provided no formal account of the effect of sampling frames.

In contrast, narrower generalization from an observed sample collected via property sampling as compared with category sampling is a core prediction of our model. With this in mind, each of the experiments in this paper included a sampling frames manipulation. In each study we tested for the predicted empirical effect of sampling fames and compared the empirical data with formal predictions from our model. However, a major strength of the censored reasoning framework is that it goes beyond an explanation of the effects of sampling frames and makes a range of additional novel predictions about how frame effects on property induction will *interact* with other aspects of the sample. As detailed below, we examine model predictions about how sample frames interact with the provision of negative evidence (Experiment 1), sample size (Experiment 2) and category base rates (Experiments 3 & 4).

2. Experiment 1: Sampling frames and the effect of negative evidence

This experiment served two goals. First, as was the case for all four experiments reported in the paper (and the additional five described in Appendix A), we tested our core prediction that generalization of a novel property will be narrower when the sample was collected via property as compared with category sampling. The paradigm for this study was adapted from Lawson and Kalish (2009). In the sampling phase, a sample of small birds was observed to have a novel property ("plaxium blood"), with the sample said to be selected via category or property sampling. Participants then judged whether the property generalized to instances of other categories that varied in similarity to the observed sample.

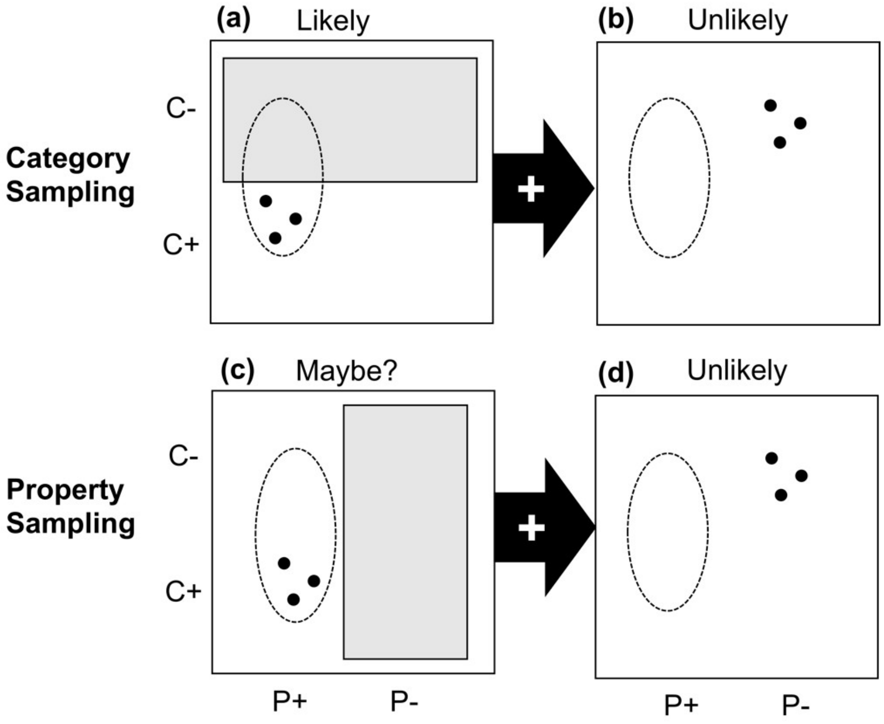


Figure 3. The effect of adding explicit negative evidence (C- P-) that escapes the censor (grey rectangle). Category sampling conditions are shown in the top row, property sampling on the bottom row. The left hand column depicts the "positive only" condition, and the right hand column shows the additional negative evidence that is introduced in the "positive + negative" condition. In all panels, the elliptical region schematically illustrates those data sets that might be expected to observe if P+ is generalizable from C+ to C-, and the markers show the data presented to participants. Note that in Experiment 1 the negative evidence was presented with a different sampling frame from that used for positive evidence in each condition (see Exp. 1 Procedure for details). To avoid making the figures unnecessarily complicated did not include a representation of this frame in panels b or d.

Our second goal was to test an additional prediction of our data censoring model by manipulating whether the observed sample contained *only* positive evidence (20 small birds all with plaxium blood) or a mixture of positive *and* explicit negative evidence (20 small birds with plaxium blood and 5 other animals *without* plaxium blood). This is illustrated in Figure 3. As described earlier, when the learner *only* sees positive evidence from the category (C+P+), it is reasonable to generalize to other categories (C-P+), but only in the category sampling scenario (panel 3a) and not under property sampling (panel 3c). The introduction of explicit negative evidence that bypasses the censor might be expected to reduce generalization in both cases (panels 3b and d), but the effect is much more pronounced for category sampling. Under property sampling, the explicit negative evidence adds little that is not already known from the implicit negative evidence. Accordingly, the theoretical prediction is that the introduction of additional negative evidence should attenuate any effect of the sampling frame, primarily by causing generalization to decrease in the category sampling condition.

2.1 Method

2.1.1 Participants.

One hundred university undergraduates (M_{AGE} = 20.89 years, SD = 3.18; 63 females) participated for course credit. Equal numbers were randomly allocated to one of four conditions (category sampling - positive only, property sampling - positive only, category sampling - positive + negative, property sampling - positive + negative). In this and all subsequent experiments, individuals gave their informed consent for experimental participation.

2.1.2 Materials and Procedure.

The evidence samples and generalization items were depicted by color pictures of birds sourced primarily from Google Images and modified using Adobe Photoshop to eliminate background. In the sampling phase we used 10 color pictures of small, sparrow-like birds, with horizontal orientation reflected to produce a total of 20 images. For the positive + negative evidence condition an additional 5 unique pictures of other animals (crow, seagull, eagle, squirrel, frog) were used in the sampling phase. For the generalization test six unique pictures were presented: a sparrow (similar but not identical to the sample instances), a pigeon, an owl, an ostrich, a mouse and a lizard. Pilot testing with participants who did not take part in the main study (N=19) examined the perceived similarity of test items to the sample pictures of small birds. Pairwise similarity ratings (1 = not very similar, 10 = very similar) confirmed that the rank ordering of perceived similarity between test and sample items was sparrow (*M* = 8.0, *SD* = 1.72), pigeon (*M* = 6.44, *SD* = 1.92), owl (*M* = 5.78, *SD* = 2.05), ostrich (*M* = 4.28, *SD* = 1.90), mouse (*M* = 2.67, *SD* = 1.65), and lizard (*M* = 2.33, *SD* = 1.61).

In the sampling phase, participants were told they were scientists studying animals on a previously unexplored island and that their task was to take samples to ascertain which animals had a novel biological property (“plaxium blood”). In the *category sampling* condition, participants were told that time and resource limitations were such that only a single category of small birds was sampled. In the *property sampling* conditions, they were told that only animals that had passed a screening test for the presence of plaxium blood were sampled (instructions were adapted from Lawson & Kalish, 2009, and are included in demonstration experiments located at http://compcogscisydney.org/exp). All participants received the same sample information.

On each of 20 sampling trials participants could click on one of 50 on-screen boxes representing all the samples taken from the island. On clicking, the box contents were revealed – showing a sample instance, depicted by a unique picture of a small bird randomly drawn from the pool of small-bird images, and a statement about whether the instance was found to have plaxium blood. The order of presentation of this information differed in category sampling and property sampling conditions. In category sampling, when a box was examined the bird picture appeared and the participant was invited to click again to discover its plaxium status. In property sampling, the order of these steps was reversed. Nevertheless, after the positive evidence trials, all participants had observed exactly the same sample information – 20 small birds with plaxium blood. Those in the positive-evidence condition then proceeded to the generalization test. Those in the positive + negative evidence condition were told that a new expedition yielded five additional sample specimens. These samples were collected under different frames to the original samples (category sampling: only animals that were not small birds sampled; property sampling: only animals that were plaxium negative). Participants again clicked on sample boxes and saw pictures of the 5 animals, none of which had plaxium blood.

In the subsequent generalization test, participants were told that 10 instances of each of a number of different types of animals from the island had now been collected. Their task was to estimate how many of these instances were likely to have plaxium based on the previously observed sample. The six generalization test instances were presented in random order and participants indicated their generalization estimate using radio buttons ranging from 0 to 10.

2.2 Results and Discussion

Property generalization scores (out of 10) for all conditions are shown in Figure 4 (see the Open Science Framework Repository (OSF) https://osf.io/j4dxm/ for raw data from this and all other experiments). These data were analyzed using a 2 (sampling frames: category, property) x 2 (evidence type: positive only, positive + negative evidence) x 6 (test item) Bayesian mixed-model analysis of variance, with repeated measures on the last factor. The analysis for this and all subsequent experiments was carried out with the JASP v0.8.6 package using Cauchy default priors (Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017). A major advantage of Bayesian approaches over many traditional analyses is that they allow for quantification of the statistical evidence in favor of *or* against the null hypothesis. The Bayes factor comparing two hypotheses is a ratio that expresses the relative probability of observing the data under one hypothesis than the other. We use the notation BF_{10} to refer to Bayes factors where $BF_{10} > 1$ indicates support for the alternative hypothesis and $BF_{10} < 1$ support for the null hypothesis. For example, $BF_{10} = 10$ indicates the data is 10 times more likely to have come from the alternative hypothesis than the null hypothesis, and $BF_{10} = 0.1$ indicates the opposite conclusion. We follow the conventions suggested by Kass and Raftery (1995) that a BF_{10} between 3-20 (0.33-0.05) represents “positive” evidence for the alternative (or null) hypotheses respectively, a BF_{10} between 21-150 (.049-.0067) represents “strong” evidence and a BF_{10} above 150 (<.0067) represents “very strong” evidence.

The analysis revealed very strong evidence of a main effect of test item, $BF_{10} > 10000$. Figure 4 shows that generalization scores were inversely related to the similarity of the test item to the evidence sample. There was also strong evidence of an effect of sampling frames, $BF_{10} = 13.63$. The figure shows that people generalized more narrowly under property sampling than under category sampling. This confirms the core prediction of our model and replicates the main finding of Lawson and Kalish (2009). The sampling frame effect did not interact with the test item factor, $BF_{10} = 0.25$.

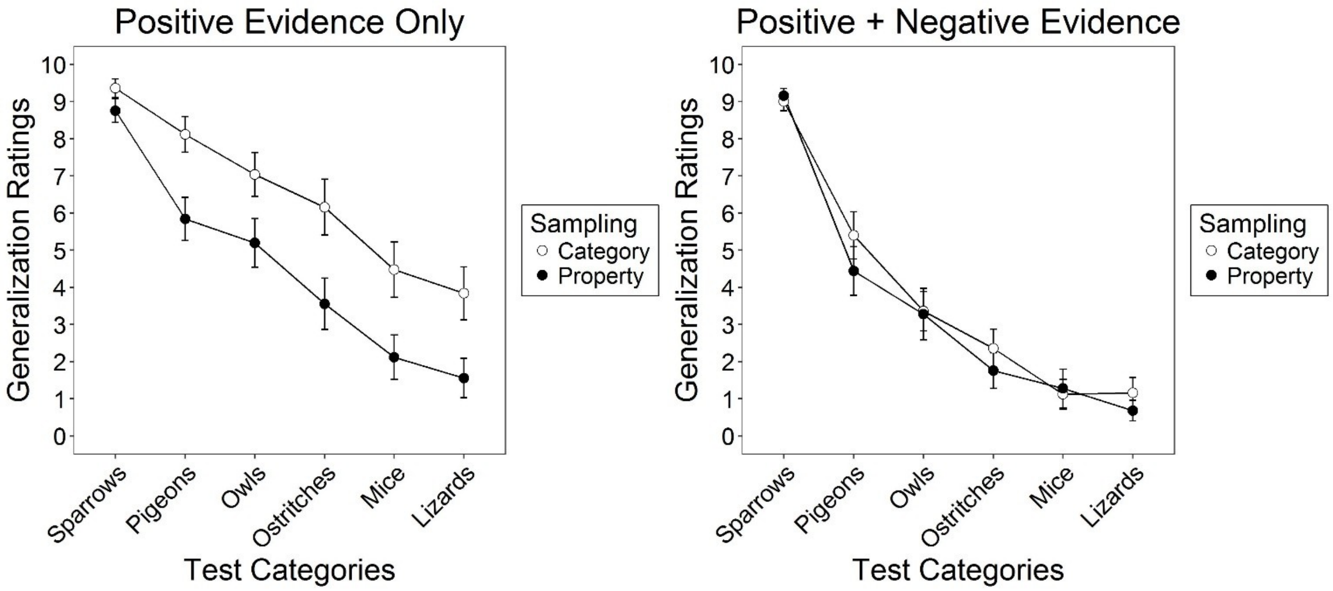


Figure 4. Mean generalization ratings for each test category. Bars represent +/-1 standard error.

There was very strong evidence of an effect of adding negative evidence, with lower generalization scores in the positive + negative than the positive only condition, BF_{10} > 10000. Evidence type also interacted with test item similarity, BF_{10} = 652.79. Figure 4 shows that the positive-only and positive + negative conditions showed similar levels of generalization to the small bird test item, but that generalization was lower in the positive + negative condition for stimuli that were less similar to the observed sample. Notably, the type of evidence observed interacted with the sampling frame, BF_{10} = 4.64. As predicted, the difference in generalization between category and property sampling was attenuated when negative evidence about large birds was added to the sample. Figure 4 shows that the explicit negative evidence had its strongest effect on generalization in category sampling. This was confirmed in post-hoc comparisons between generalization scores in the positive evidence only and positive + negative conditions, where evidence of a difference was found for category sampling (BF_{10} > 10000) but not property sampling (BF_{10} = 1.38).

This study tested two predictions of our censored sampling model. First, we showed that providing a property sampling frame for a given evidence sample led to narrower property generalization than when the same sample was obtained via category sampling. This replicates the main finding of Lawson and Kalish (2009). However, unlike that earlier work, we have provided a formal model that explains this effect.

The second and entirely novel prediction of our model was that generalization following category sampling would be impacted more by adding explicit negative evidence than generalization following property sampling. This prediction was also confirmed. Providing evidence about instances that did not have the property narrowed generalization following category sampling because this evidence ruled out hypotheses about property extension (e.g., that the property generalized to large birds) that were still viable when only the positive instances were observed.[1]

3. Experiment 2: The effect of sample size

Experiment 1 found evidence for a sampling frames effect, and demonstrated one way in which the effect can be manipulated: adding explicit negative evidence selectively influences generalization under category sampling. In this experiment we considered a manipulation expected to selectively influence generalization in the property sampling condition: sample size. According to our theory of the frames effect, a failure to observe the C-P+ case is more informative under a property sampling frame than under a category sampling frame. However, the strength of this effect depends on how many instances have been sampled. Under property sampling, observing that the first few instances known to share a property all belong to a single category may be attributed to chance, preserving the belief that the property may generalize to unobserved categories (see Figure 5 for a schematic illustration). The absence of certain types of category members in a *large* sample selected on the basis of property (like those used in Experiment 1), seems more conspicuous. It licenses the inference that the property does not extend very far beyond the observed sample. In other words, for small samples we should see little difference in property generalization between category and property sampling conditions. With increasing sample size, generalization under property sampling should diverge from category sampling – with property generalization increasingly restricted to items similar to the sample.

This experiment tested the prediction about the modulating effects of sample size on sampling frames by asking participants in property and category sampling conditions to make repeated generalization judgments after observing samples of 2, 6, and 12 instances in a within-subjects design.[2] In this experiment we also changed the cover story, switching to a more novel conceptual domain (learning about the properties of rocks on the planet Sodor). There were two reasons for this. First, we wanted to test the generality of the sampling frames effect across cover stories. Second, we wanted to simplify the conceptual space for generalization. In Experiment 1 we examined property generalization across complex animal categories that differ on a range of perceptual and conceptual dimensions. While studying generalization across such naturalistic categories is interesting, modeling the relevant similarity space and factoring this into predictions about property generalization is non-trivial (cf. Nosofsky, Sanders, & McDaniel, 2018). Because modeling of similarity spaces was not the main goal of the current work, in this and subsequent experiments we used a generalization test set where similarity to the observed sample varied across a single salient dimension (i.e., size). As detailed below, we also made some more minor changes to the experimental procedure to reduce the length of the sampling phase and make the task more amenable to on-line data collection.

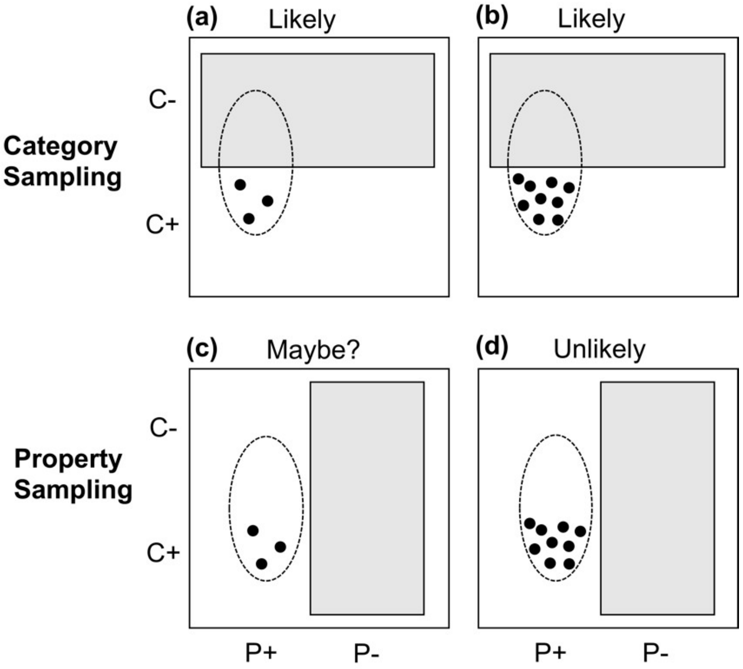


Figure 5. The predicted interaction between sample size and sampling frame. Under category sampling, observing additional C+P+ evidence does not lead to very much belief revision: regardless of whether few observations (panel a) or many (panel b) are observed, the data are exactly what one would expect to observe if the property is shared with non-category members C-. Under property sampling, however, a different pattern is seen: if the C-P+ case exists, one should expect to eventually encounter it in a property sampling scenario so as more examples of the C+P+ case are observed (panels c and d), the strength of evidence against generalization increases.

3.1 Method

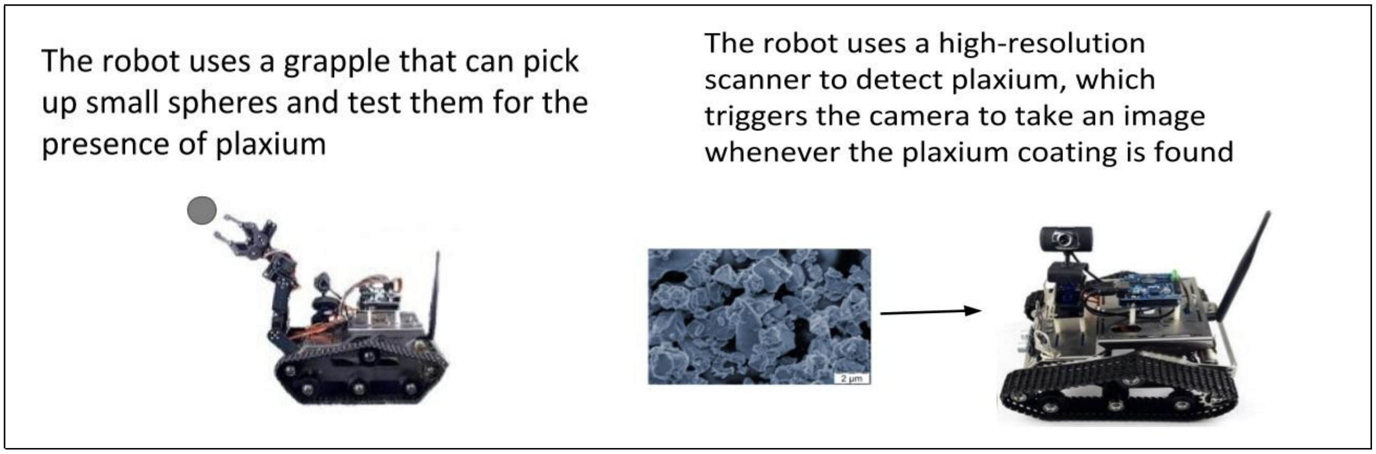
3.1.1 Participants.

225 on-line participants were recruited using Amazon Mechanical Turk (AMT). In this and subsequent studies, all participants were from the USA and were AMT workers with a minimum approval rating of 95% for previous AMT work (M_{AGE} = 34.92 years, SD = 11.65; 112 females). They were paid $1.25 US on task completion. Participants were randomly allocated to either the category (n=114) or property sampling condition (n=111).

3.1.2 Procedure.

The task structure was similar to Experiment 1. Participants first saw identical samples of evidence about objects that had a novel property under either category or property sampling instructions, and then inferred whether the property generalized to objects that varied in similarity to the sample. However, the cover story, stimuli and some aspects of the procedure were modified.

Participants were told that they were scientists in the future studying the newly discovered planet Sodor. They were informed via text and pictorial illustrations that rocks on Sodor were circular and varied in size. Their task was to use a robot on the planet’s surface to discover which rocks contained the valuable substance “plaxium”. Before commencement of the sampling phase, those in the category and property sampling conditions were given different explanations of constraints on the sampling process (see Figure 6). In category sampling, only small rocks were sampled because only these would fit into the robot’s small collecting claw. In property sampling, the rocks sampled were the first to show a positive result when photographed using a plaxium-sensitive camera. Participants were not permitted to proceed to the sampling phase until they achieved a perfect score a 3-item multiple choice-test that assessed comprehension of the instructions. If this test was failed the participant was returned to the instruction screens. All participants passed this comprehension test within four attempts.



The robot uses a grapple that can pick up small spheres and test them for the presence of plaxium

The robot uses a high-resolution scanner to detect plaxium, which triggers the camera to take an image whenever the plaxium coating is found

Figure 6. Screen shots of instructions used for category sampling (left frame) and property sampling (right frame).

On each sampling phase trial participants clicked a button to reveal a rock sample. The sample item then appeared with the message “plaxium detected”. After each item appeared, participants had to wait for 3 s before collecting the next sample. The pictures of each sample accumulated on the screen. All participants viewed the same sample of rocks, which were at the "small" end of the size dimension (diameters ranging from 0.4 cm - 0.6 cm) with presentation order of items randomized for each participant.

At three points in sampling (after observing 2, 6 and 12 rocks), the collection of new samples was paused and participants were probed for inferences about whether plaxium generalized to seven test rocks varying in size (diameters in cm: 0.4, 0.6, 1.0, 1.4, 1.8, 2.2, 2.6). Participants had to rate the likelihood that each test item had plaxium (1 = definitely does not, 10 = definitely does). Test items were presented in random order. The two smallest test rocks (R1, R2) were the same size as rocks observed in the sample.

3.2 Results and Discussion

Generalization ratings are shown in Figure 7. These data were analyzed using a 2 (sampling frames: category, property) x 3 (sample size: 2, 6, 12) x 7 (test item) Bayesian mixed-model analysis of variance, with repeated measures on the last two factors.

Property generalization decreased as the size of the test rocks increased (i.e. as they became less similar to the observed sample), BF_{10} > 10000. Generalization ratings averaged across test items were higher overall for smaller than for larger samples, BF_{10} > 10000. Overall generalization was also stronger following category than property sampling, BF_{10} > 10000. The latter finding replicates the key effect of sampling frames from Experiment 1, showing that this effect extends to our new cover story and stimuli.

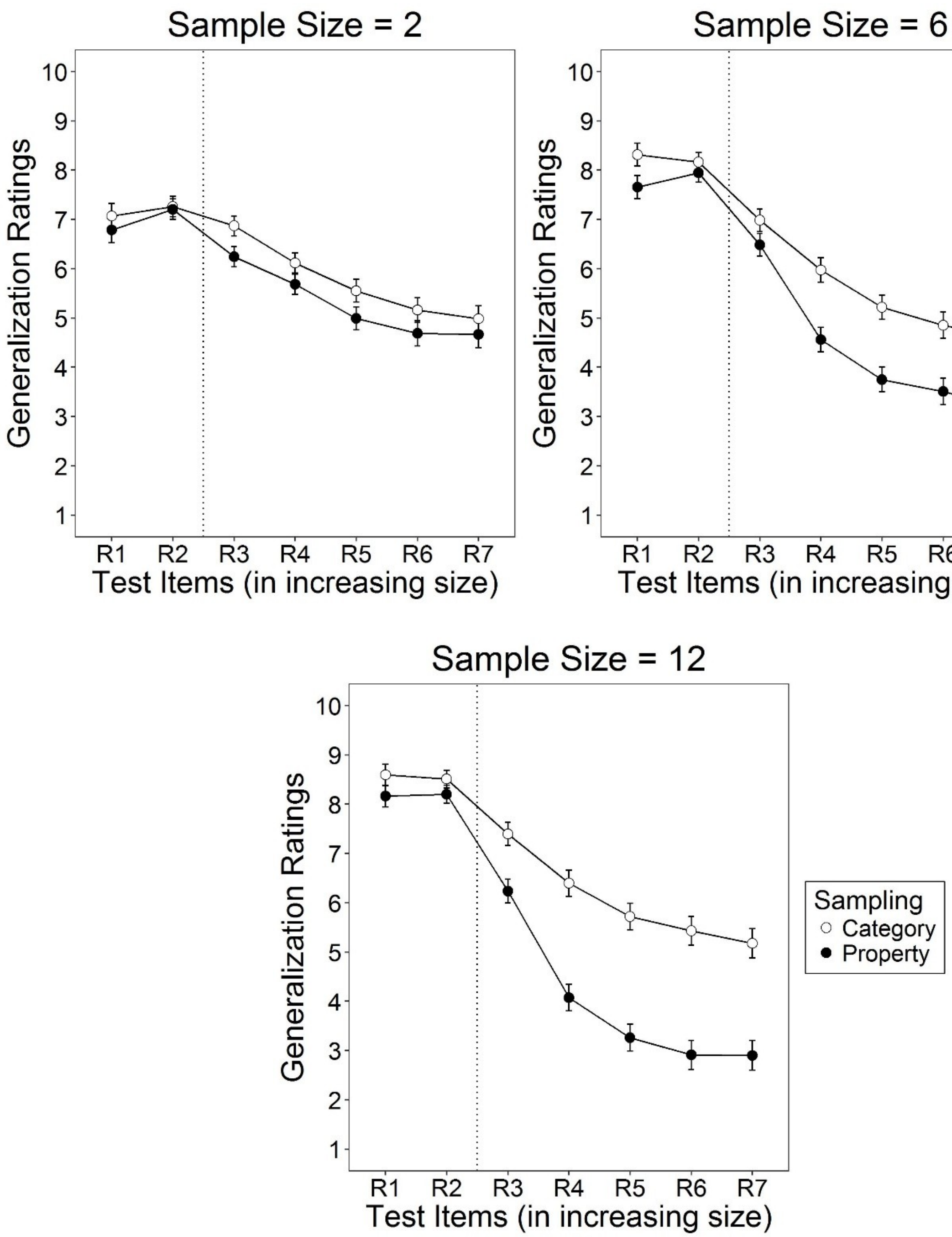


Figure 7. *Mean generalization ratings for each Sodor rock (R1-R7) based on sample sizes of 2, 6 and 12 items. Bars represent +/-1 standard error. Items to the left of the dotted vertical line were identical to those presented during sampling. Items to the right were novel.*

The effects of both sampling frames and sample size varied across test items, with BF_{10} > 10000 for each two-way interaction. Both of these factors had more of an impact on generalization to test rocks that were larger than those observed during sampling (i.e. R4-R7). Crucially, the effect of sampling frames was modulated by sample size, BF_{10} > 10000. Figure 7 shows that after observing only two instances, people in both property and category sampling conditions responded conservatively, giving generalization ratings in the mid-range of the scale. As sample size increased, there was increasing differentiation in generalization following property or category sampling. This effect was especially pronounced for the largest test items (i.e., those most dissimilar to the sample instances), BF_{10} = 7.30. Follow-up tests confirmed that for the smallest sample there was indeterminate evidence of an effect of sampling frame on generalization, BF_{10} = 0.47. However, there was robust evidence of a difference between property and category sampling frame at sample sizes six, BF_{10} = 2723.70, and twelve, BF_{10} > 10000.

The moderating effect of sample size on the sampling frames manipulation is consistent with our model of inference. Those in the property and category sampling conditions responded in similar ways to an absence of large rocks from the smallest sample. Presumably this absence was simply attributed to chance. As the size of the sample increased however, this absence was viewed by those in the property sampling condition as cumulative evidence that the target property was restricted to small Sodor rocks.

4. Experiment 3: The role of base rates

The first two experiments showed that people are sensitive to sampling frames, with the magnitude of the frames effect changing in sensible and predictable ways when we introduced negative evidence (Experiment 1) or changed the sample size (Experiment 2). This experiment tests a third factor that should interact with sampling frames during the inference process – the *base rates* of different types of categories within a population.

In previous studies those in the property frames condition who observed a large sample from a given (target) category but no instances of other categories inferred that the novel property did not generalize very far beyond the sample. But this inference is only licensed if one assumes that instances from non-target categories *could* have been observed during sampling. If instances of unobserved categories were rare in the population, their absence from the sample is less informative for property generalization. This is schematically illustrated in Figure 8.

To test this prediction, we manipulated both sampling frames and category base rates. The general design was similar to Experiment 2 except that only a single sample size (9 items) was used and people were informed about the relative base rates of the observed and unobserved categories. In the C+ rare/C- common condition (hereafter C- common for the sake of brevity), the members of the observed target category (small rocks) were said to be rare and members of the unobserved category (large rocks) common. The C+ common /C- rare condition (hereafter C- rare) reversed these base rates with observed target category members common and non-targets rare.

The data from the property sampling conditions in the previous studies suggest that, in the absence of base rate information, people generally assume that members of non-target categories could have been observed during sampling. Hence, we expected that the effects of sample frames on inferences in the C- common condition would be similar to those observed previously. In the C- rare condition however, the absence of larger rocks in the property frame sample can be attributed to their low base rate rather than their lack of plaxium. Hence, we should see smaller differences in property generalization between category and property frame conditions in the C- rare condition.

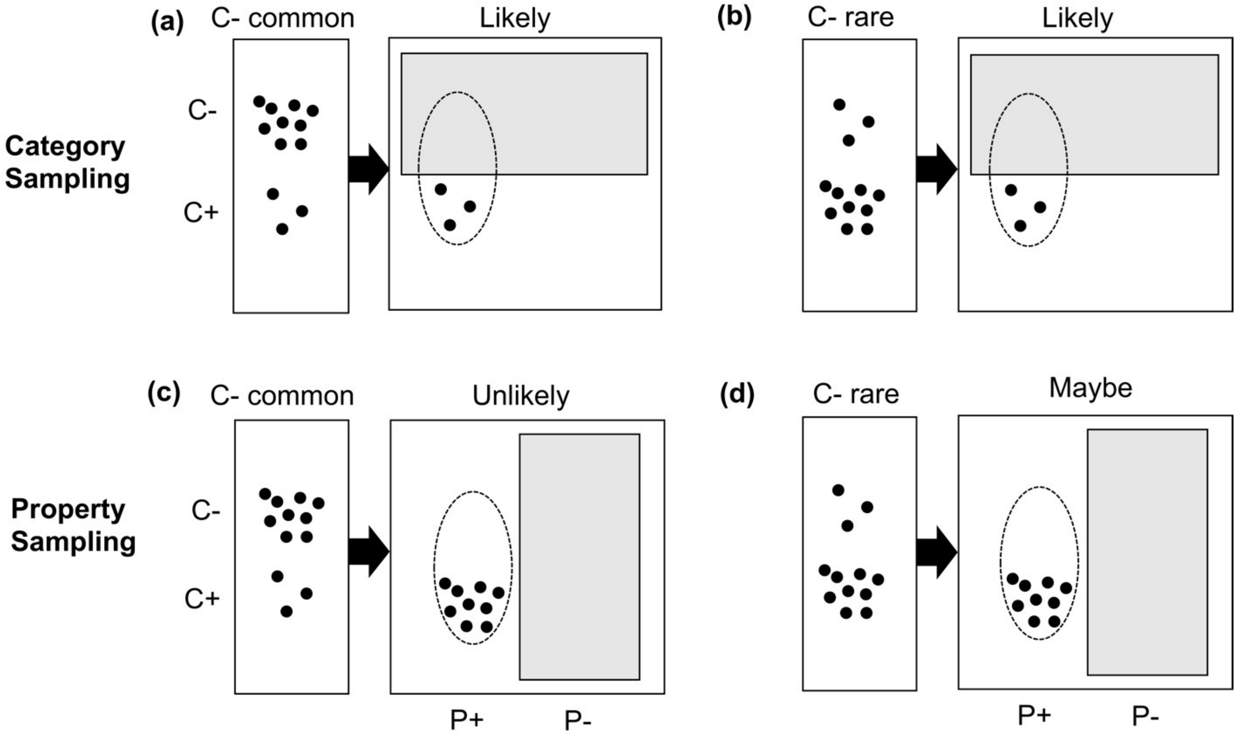


Figure 8. *The predicted effect of base rates in category sampling and property sampling.*

These predictions assume that people are sensitive to sample base rates and will combine them with other sample information such as frame constraints, when drawing inferences. This may seem at odds with findings that people frequently ignore base rates when making probabilistic judgments (see Meder & Gigerenzer, 2014 for a review), especially when base rates are described in words and/or statistics rather than being experienced in trial-by-trial sampling (Hawkins, Hayes, Donkin, Newell, Pasqualino, & Newell, 2015; Hogarth & Soyer, 2011). Indeed, some have cited base rate neglect as evidence that reasoners are “naïve statisticians” (e.g., Juslin, Winman, & Hansson, 2007) or are generally “myopic” about the implications of base rates and other parameters that affect sample composition and generation (e.g., Fiedler, 2012; Kahneman, 2011).

In some respects, this pessimistic view seems at odds with our sampling frame results. The previous studies show that reasoners faced with identical samples draw very different inferences depending on how the sample was selected. In this respect they showed no myopia in their understanding of how sampling processes affect inference. It remains to be seen whether participants are capable of factoring *both* frame constraints and base rates into their inferences.

4.1 Method

4.1.1 Participants.

789 on-line participants were recruited using Amazon Mechanical Turk and were paid $1.25 US on task completion (M_{AGE} = 35.72 years, SD = 11.13; 328 females). The large sample was motivated by pilot work suggesting that if a moderating effect of base rates on sampling frames was found, it would likely be small (see Appendix A). Participants were randomly allocated to one of four conditions: category sampling, C- common (n = 211), category sampling, C- rare (n = 194), property sampling, C- common (n = 203), property sampling, C- rare (n = 181).

4.1.2 Procedure.

The procedure was similar to Experiment 2, except for the following. Prior to the sampling phase participants were shown two screens that provided information about the relative base rates of large and small rocks on Sodor. In the C- common condition, participants were shown a picture of six “large” circles (diameter > 1.0 cm) and one “small” circle (diameter = 0.4 cm), arranged in a random fashion, with text stating that “Most Sodor rocks are large. Small Sodor rocks do exist but are very rare”. In the C- rare condition, the base rate pictures and text were reversed with one large small and six small circles shown. To ensure correct encoding of base rates, the multiple choice test that preceded sampling included a base rate question that had to be answered correctly before the participant was allowed to proceed.

The sampling phase followed the presentation of base rate information. The category and property frames instructions were the same as in Experiment 2. The sampling phase trials proceeded in a similar manner to the previous study except that all participants saw a sample of nine small Sodor rocks, and only made one set of generalization ratings after all samples were viewed.

4.2 Results and Discussion

Generalization ratings are shown in Figure 9. These data were analyzed using a 2 (sampling frames: category, property) x 2 (base rate: C- common, C- rare) x 7 (test item) Bayesian mixed-model analysis of variance, with repeated measures on the last factor. We again found that generalization decreased as the size of test items increased, BF_{10} > 10000. We also again found very strong evidence for a main effect of sampling frames, with less generalization to test items following property sampling than category sampling, BF_{10} > 10000. The sample frames effect did not vary across the range of test items, BF_{10} = 0.0009. Overall, generalization ratings were higher in the C- rare base rate condition than in the C- common condition, BF_{10} = 34.91. As indicated by an interaction between the base rate and test items factors, BF_{10} = 127.28, the higher generalization ratings in the C- rare condition were concentrated among the novel, larger test items (i.e. R3-R7). However, no clear evidence was found for the predicted interaction between sample frames and base rate, BF_{10} = 0.53.

This experiment again replicated the sampling frames effect with property sampling leading to more restricted generalization than category sampling. Participants were also more likely to generalize the novel property to large Sodor rocks when these types of rocks were rare (C- rare) than when they were common (C- common). However, contrary to predictions we did not find evidence that the sampling frames effect was modulated by the relative base rates of observed and unobserved evidence.

Moreover, three other variations of this experiment using very similar designs (B1, B2 and B3 in Appendix A) found essentially the same thing.

An overall effect of base rates on generalization to novel test items indicates that participants did encode the relevant base rates and used them in their inferences about property generation. This suggests that people may not have been entirely "myopic" about the importance of sample base rates for property inference (Fiedler, 2012), but nevertheless, it appears that those in the property sampling C- rare condition did not see the base rates as an alternative explanation for why only small rocks were observed in the sample.

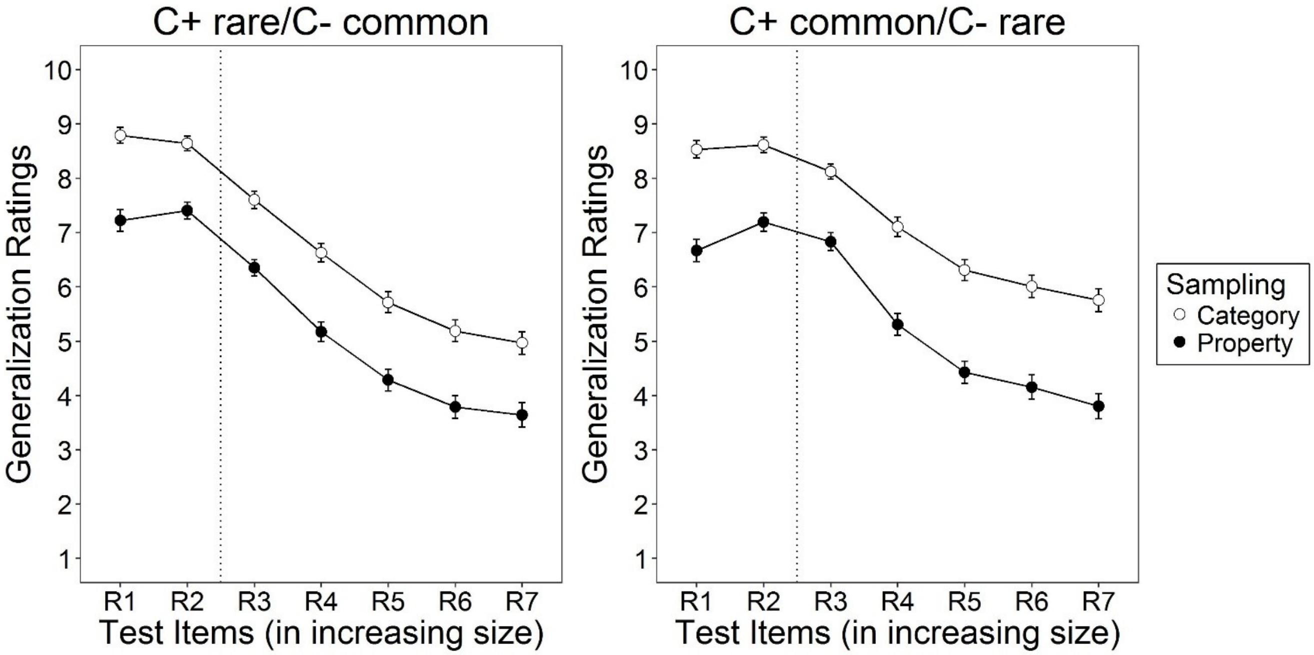


Figure 9. Mean generalization ratings for each Sodor rock (R1-R7). Bars represent +/-1 standard error. Items to the left of the dotted vertical line were identical to those presented during sampling. Items to the right were novel.

5. Experiment 4: Base rate effects revisited

The absence of an interaction between base rate and sampling frame in Experiment 3 is at odds with the predictions of the censored reasoning model, though consistent with other studies demonstrating base rate neglect in statistical reasoning (e.g., Chun & Kruglanski, 2006; Hayes, Hawkins & Newell, 2016; Koehler, 1996). One factor that can affect the integration of base rates with other relevant problem components is their relative salience. The extent to which people use base rate as opposed to individuating details in lawyer-engineer problems for example, is related to the amount of detail provided about each type of information (cf. Koehler, 1996; Welsh & Navarro, 2012). In the typical problem where base rate neglect is observed, base rates are mentioned only briefly while a rich description of individual features is provided. However, when the description of base rates is more extensive, perhaps unsurprisingly, they are more likely to be combined with other statistical information in subsequent judgments (e.g., Chun & Kruglanski, 2006).

One possibility is that the design used in Experiment 3 (and studies B1-B3 in Appendix A) used a detailed instruction set and trial-by-trial learning in order to instantiate the sampling frames, but did not provide a similarly "salient" method for describing category base rates. To address this – and to see if it is possible to obtain the predicted interaction in at least some experimental designs – Experiment 4 adopts a more "heavy handed" approach, using a stronger base rate manipulation.

5.2.1 Method

5.1.1 Participants.

Participants were recruited using AMT, with the same payment and inclusion conditions as previous studies (M_{AGE} = 35.68 years, SD = 10.81; 190 females). Random allocation resulted in the following cells: category sampling, C- common (n = 110), category sampling, C- rare (n = 107), property sampling, C- common (n = 111), property sampling, C- rare (n = 95).

5.1.2 Procedure.

This was identical to the previous study except for changes to the way base rate information was presented. First, the base rate difference between small/large Sodor rocks was more extreme (ratio of approximately 25:1 as compared to 6:1 in the earlier experiment). Second, an additional instruction screen was used to illustrate the base rate discrepancy. This presented the relative numbers of small/large Sodor rocks in two vertical columns presented side-by-side (see Figure 10).

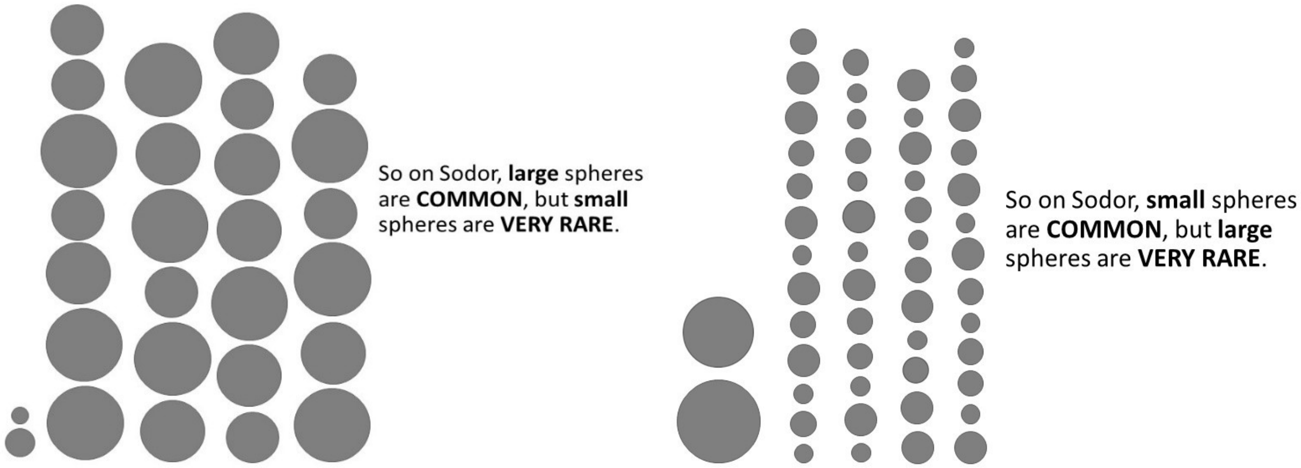

So on Sodor, **large** spheres are **COMMON**, but **small** spheres are **VERY RARE**.
So on Sodor, **small** spheres are **COMMON**, but **large** spheres are **VERY RARE**.

Figure 10. Screens used to illustrate base rates in the C+ rare/ C- common condition (left) and C+ common/ C- rare condition (right)

5.2 Results and Discussion

Property generalization ratings are shown in Figure 11. A Bayesian mixed-model analysis of variance replicated the key main effects of test items, sampling frames, and base rates from the previous study, all, BF_{10}*'s* > 10000. There was also evidence of a two way-interaction between frames and test items, BF_{10} = 343.39, and between base rates and test items, BF_{10} > 10000. However, in this case we also found strong evidence for a two-way interaction between sampling frames and base rates, BF_{10} = 24.02, and a three-way interaction between sampling frame, base rate and test items, BF_{10} = 165.25. Figure 11 shows that, in line with our model predictions, i) generalization to items that were part of the observed sample (R1-R2) was lower in the property sampling, C- rare condition than in other conditions, and that

ii) the differences between category and property sampling in generalization to the largest new items (R5-R7) were attenuated in the C- rare as compared to the C- common condition. The latter effect was further confirmed by Bayesian t-tests comparing category and property sampling for each test item in each base rate condition. For the C- rare condition, there was strong evidence that category sampling led to higher generalization ratings than property sampling for R1, R2, R3 and R4 (all BF_{10}'s >250), but no clear evidence of a frames effect for the largest (i.e. most dissimilar) test items, R5, R6 and R7 (BF_{10}'s = 0.92-0.28). By comparison, in the C- common condition, category sampling led to higher generalization ratings than property sampling for *all* test items (BF_{10}'s = 8379.08 – 11.27).

As predicted, base rate information had a profound effect on the way that people drew inferences from property sampling. When small rocks were common and large rocks were rare this provided an alternative explanation for the composition of the property sample (aside from sharing plaxium). The base rate information preserves the hypothesis that large rocks could have plaxium but were unlikely to be observed in the property sample. As well as providing further support for our Bayesian account, these results show that people are capable of integrating information about two sampling parameters (frames and base rates) into their property inferences.

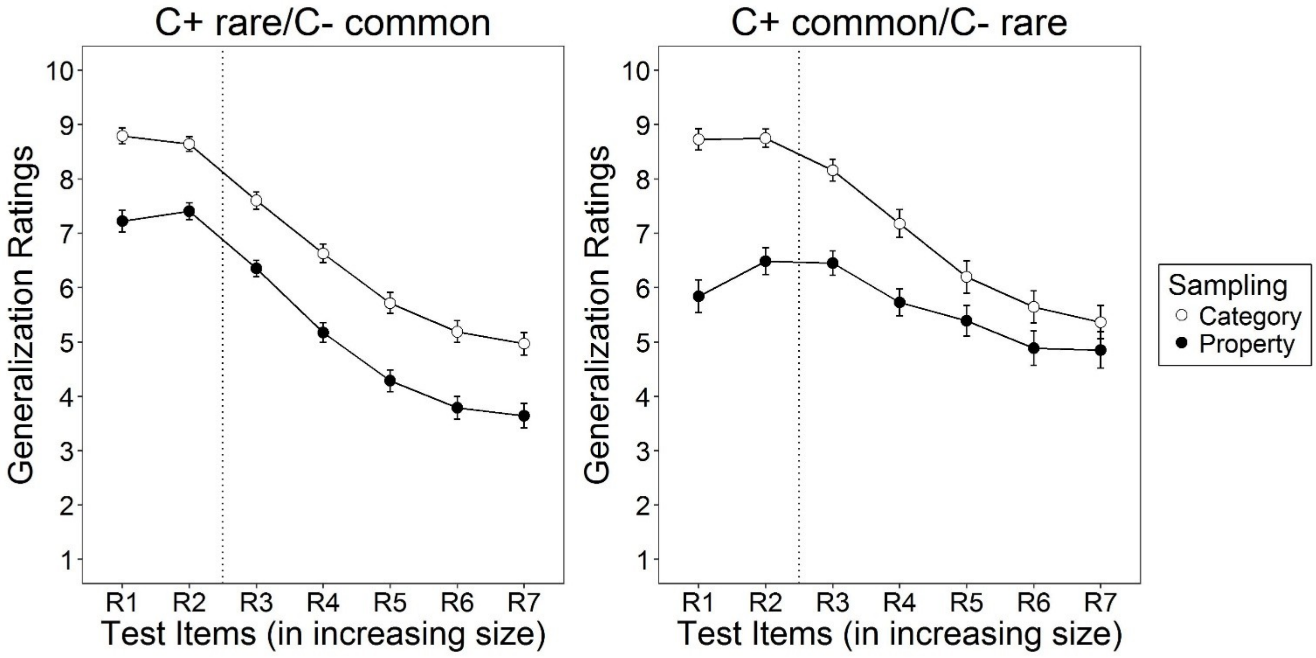


Figure 11. *Mean generalization ratings for each Sodor rock (R1-R7). Bars represent +/-1 standard error. Items to the left of the dotted vertical line were identical to those presented during sampling. Items to the right were novel.*

6. A new computational model of inference with censored samples

So far, we have mainly motivated our experiments by relying on schematic visual depictions (Figures 2, 3, 5 and 8) to link our theoretical framework (Figure 1) to the experimental work. In this section, we introduce a specific instantiation of this framework, which we will then compare to the experimental data. Our model is consistent with the Bayesian approach to inductive reasoning (see e.g., Tenenbaum & Griffiths 2001) but extends it in two novel directions. The first and most important innovation – which ties to the central topic of the paper – is the introduction of an explicit representation of the sampling frame (denoted *S*) to describe the effect of *mechanistic* constraints on sampling processes. In previous theoretical work on this topic, sampling processes are implicit insofar as they place constraints on the likelihood function – for instance, the probability of the observations $P(d|h)$ is might be different under "strong sampling" versus "weak sampling", as per Tenenbaum and Griffiths (2001), or perhaps different under "helpful sampling" versus "random sampling" as per Shafto, Goodman and Griffiths (2014). None of these existing sampling rules precisely matches the category sampling or property sampling scenarios we consider in our experiments, but they are easily represented as different kinds of censoring mechanisms *S*. Under category sampling, the selection mechanism *S* constrains the observation to be a small bird, and the likelihood function describes the probability that *a small bird has plaxium blood*. Under property sampling, this is reversed: the selection mechanism restricts the observation to be an animal with plaxium blood, and the likelihood describes the conditional probability that *an animal with plaxium blood is a small bird*.

The second innovation is the implementation of a function learning mechanism in which each hypothesis *h* maps to a smooth generalization function (denoted *f*, see below) which allows the model to make predictions about people's inferences over a "continuous" generalization space (e.g., animals that vary over a continuous similarity space, Sodor rocks of increasing size). Previous Bayesian models of inference (e.g., Hayes et al., in press; Hendrickson et al., under review; Navarro et al., 2012; Ransom et al., 2016; Tenenbaum & Griffiths, 2001) have generally relied on the idea of "consequential regions" idea introduced by Shepard (1987; see also Soto, Gershman & Niv 2014). In essence, this involves determining whether a novel test stimulus is likely to belong to the same region in psychological space as a familiar stimulus. Instead, we adopt a novel approach in which the learner's goal is to infer a continuous-valued *function* defined over the stimulus space. That is, the learner aims to learn the *probability* that a particular type of entity (e.g., small bird, small Sodor rock) possesses a property (e.g., plaxium). The consequential region approach is a special case of the function learning perspective, in which each hypothesis (region) is a function *f*(x) that is constrained to have values of 0 or 1. In effect, the learner assumes that the category of items that possesses a property has hard boundaries (it either does or it doesn't have the property). By assuming instead that the function *f*(x) is smooth and continuous, our perspective allows for probabilistic inferences about graded categories in a more natural way and avoids the need to impose strict a priori boundaries between categories.[3] Although previously applied in Bayesian approaches to human function learning (e.g., Lucas, Griffiths, Williams & Kalish 2015) and classical conditioning (Lee, Lovibond, Hayes, & Navarro, in press), to our knowledge this is the first time a function learning mechanism has been incorporated into a Bayesian model of reasoning.

6.1 Model details

Inspired by work in the Bayesian function learning literature (Griffiths, Lucas, Williams, & Kalish 2009; Lucas, et al., 2015), we define the computational problem for property induction tasks as follows. We assume that the reasoner's goal is to learn a smooth function f that specifies the logit of the probability $\phi(x)=f(x)/(1-f(x))$ that any given entity$x$ has plaxium blood (or similar). We place a Gaussian process prior $P(f)$ over this function (Rasmussen & Williams, 2006), which ensures that similar entities have similar probabilities, but is otherwise unconstrained.

The Gaussian process (GP) provides a method for specifying priors over smooth functions (see Rasmussen & Williams 2006; Schulz, Speekenbrink & Krause 2018). The goal is to infer a prediction function *f*: $R^m \longrightarrow R$that maps every possible stimulus *x* (presumed to be represented as real-valued vector of length *m*) onto some subjective notion of inductive strength *y* = *f(x)*. The function *f* is defined over the entire stimulus space, but is measured a finite subset of points *x* = $(x_1, \ldots, x_n)$. The key idea in Gaussian processes is that for every possible finite subset of input variables *x*, the joint distribution over the corresponding output variables *y* = $(y_1, \ldots, y_n)$ is a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ. This prior is denoted:

$$f(x) \sim \mathrm{GP}(\mu, \Sigma) \qquad (2)$$

In the function learning context, the inductive problem facing the human learner is similar in nature to regression problems in statistics, and this prior is sensible (e.g., Lucas et al 2016). For property induction tasks, however, participants are implicitly solving a binary classification problem, deciding on the probability that a stimulus possesses a particular property. To do so, the function *f* is mapped onto the unit interval by passing it through a logistic function,

$$\phi(x)=\frac{1}{1+\exp(-f(x))}(3)$$

producing a function $R\rightarrow[0,1]$ whose values can be interpreted as probabilities. After observing a set of stimuli and their properties (e.g., small birds with plaxium blood), the learner updates the prior a posterior distribution over possible inductive generalization functions ϕ. The curves plotted in Figure 12 show the posterior mean values for ϕ.

6.1.1 Setting model parameters

Under the GP prior, the prior mean of the function *f* at points *x* is given by the mean vector μ, which we fix at 0 for all values of *x*, which in turn implies that the prior mean for the generalization function ϕ is 0.5 for all possible stimuli. This is arguably plausible for property induction experiments involving “blank” predicates, but in other situations where people bring real world knowledge to the task a more flexible approach may be required.

The covariance matrix $\Sigma = [\sigma_{ij}]$ governs the smoothness of the function *f*, and is defined in terms of a kernel function $\sigma_{ij} = K(x_i, x_k)$ that specifies how correlated the values of *f* are as a function of the stimulus properties x_i and x_j. There are a variety of kernel functions used in the statistical learning literature, but for the purposes of the current paper we focus on radial basis functions, in which the similarity between items x_i and x_j depends only on the distance between them, $d_{ij} = |x_i - x_j|$ in an appropriately formulated psychological space. Insofar as the inductive generalization problems considered here are a form of similarity-based generalization, the radial basis function allows us to describe the problem solely in terms of the psychological distance between items (see Jäkel, Schölkopf & Wichmann, 2009). Thus, while the unidimensional stimulus representation shown in Figure 12 seems reasonable as a way of representing the Sodor sphere in Experiments 2-4, the formalism used here is sufficiently general that it also applies to the animals task (Experiment 1), so long as the distances between items in the mental representation are reasonably closely approximated by the set of distances between points on a line. No implication of psychological unidimensionality is required. In all simulations, we make the simplifying assumption that the test items *x* are spread evenly across the stimulus space, with values *x* = 1, 2, ... , n. The specific kernel function we use is

$$K(x_i,\ x_j) = \tau^2 \exp(-\rho d_{ij}^{\ 2}) \qquad (4)$$

In this expression, *τ* describes a baseline correlation between pairs of items, and *ρ* governs the rate with which this correlation decays as function of distance. The elements of the covariance matrix Σ are then given by:

$$\sigma_{ij} = \begin{cases} K\left(x_i, x_j\right) \text{if } i \neq j \\ K\left(x_i, x_j\right) + \sigma 2 \text{ if } i = j \end{cases} (5)$$

where *σ* describes the inherent noise in the data. Parameter setting was done by hand and fixed at *σ* = .5, *τ* = 1.5 and *ρ* = .1 for all modelling exercises. That said, it is important to note that an exploration of the parameter space of the model (cf. Navarro, Pitt & Myung 2004; Pitt, Kim, Navarro & Myung 2006) revealed that the success of the model predictions does not depend on the selection of these particular parameter values (see Appendix B).

6.1.2 Category sampling and property sampling

In almost all cases, the only data available to the learner were property-positive exemplars of a target category. Under property sampling, the censoring function *S*(*x*) only allows property-positive items to be included in the sampling frame, whereas under category sampling *S*(*x*) only admits category members. We implement these in the following way. Under the hypothesis *h* that *ϕ*(x) describes the true probability that a member of category *x* possesses the property, the probability of observing a property-positive example of category *x* is simply $P(d|h) = \phi(x)$, and the probability of several such examples is the product of their individual probabilities. Under property sampling, the situation is a little different. Let $\theta(x)$ denote the prior probability that a randomly sampled exemplar belongs to category *x* (i.e., the base rate). Then the probability of observing a member of the category that is not censored out (i.e., possesses the property) is the product $\theta(x)\phi(\mathrm{x})$ so the likelihood function for a single observation becomes

$$P(d|h) \propto \theta(x)\phi(x) \qquad (6)$$

where, except as specified below, for a finite set of categories *x* we place a symmetric Dirichlet prior over the category base rates, $\theta(x) \sim \mathrm{Dirichlet}\ (\alpha)$. In our model simulation we set $\alpha = .35$.

6.1.3 Describing the role of moderator variables

The experiments explore three moderating variables: sample size, base rates, and explicit negative evidence. To capture the sample size manipulation we vary *n* so that it matches the true number of observations presented to participants. For the base rate manipulations, we alter the Dirichlet prior over base rates so that the value of α for the target category (e.g., small spheres) is either very high or very low: we used α = .01 for the rare target condition and α = 20 for the common target condition.

In the explicit negative evidence condition, participants were told that the negative evidence items were selected using an “inverted” version of the sampling frame: for property sampling, they were told that animals were selected because they were plaxium negative, whereas in category sampling the items were selected because they were not small birds. Accordingly, for property sampling the likelihood of the negative evidence items is

$$P(d|h) \propto (1 - \theta(x))\phi(x) \qquad (7)$$

Under category sampling there is some ambiguity as to whether sampling “other animals” implies that the set of animals was fixed by the sampling frame or whether they were sampled at random conditional on not being small birds. For simplicity, we assume the former, but both yield the same predicted behaviors. Hence, the probability of observing a plaxium negative example is

$$P(d|h) = 1 - \phi(x) \qquad (8)$$

6.2 Testing the model predictions against the data

How closely does the Gaussian process Bayesian model mirror human performance in the experiments? Overall, as shown in Figures 12 and 13, the model performs very similarly to human participants for Experiments 1, 2, and 4 (see Appendix C for modeling of Experiment 3). Not only does it reproduce the core effect of sampling frames, it also reproduces almost all of the qualitative effects found in the experiments: adding explicit negative evidence attenuates the effect (left panels), increasing sample size exaggerates it (middle panels) and changing the base rate affects property sampling but not category sampling (right panels). In some instances, it performs surprisingly well at capturing subtle details of the data. For example, when sample size increases this affects generalization in both property and category sampling, but the qualitative pattern of this change is different. For category sampling, the bulk of the effect occurs for the target category and highly similar categories (i.e., the curves shift up on the left-hand side), with no effect on the distant categories. Under property sampling, there is a more modest sized effect, but it occurs as a crossover that shifts ratings up for very similar items and down for very dissimilar items. The Gaussian process model reproduces this pattern.

That said, there are also a number of places where the model and human inferences differ. The most noteworthy are (1) the GP model is less willing to endorse the target category under property sampling (most notably in Experiment 1), because it allows for the possibility that there are plaxium-negative category members that were censored out during sampling. People appear less likely to consider this possibility. Additionally, (2) the GP model predicts a genuine null effect of base rate under category sampling, whereas the Experiment 4 data merely show an attenuated one. Nevertheless, minor prediction failures notwithstanding, the model performs admirably well with very little model fitting (see Appendix B), correlating above *r* = 0.9 with the human data in all three experiments, as shown in Figure 13 (the corresponding correlation for Experiment 3 was *r* = 0.84).

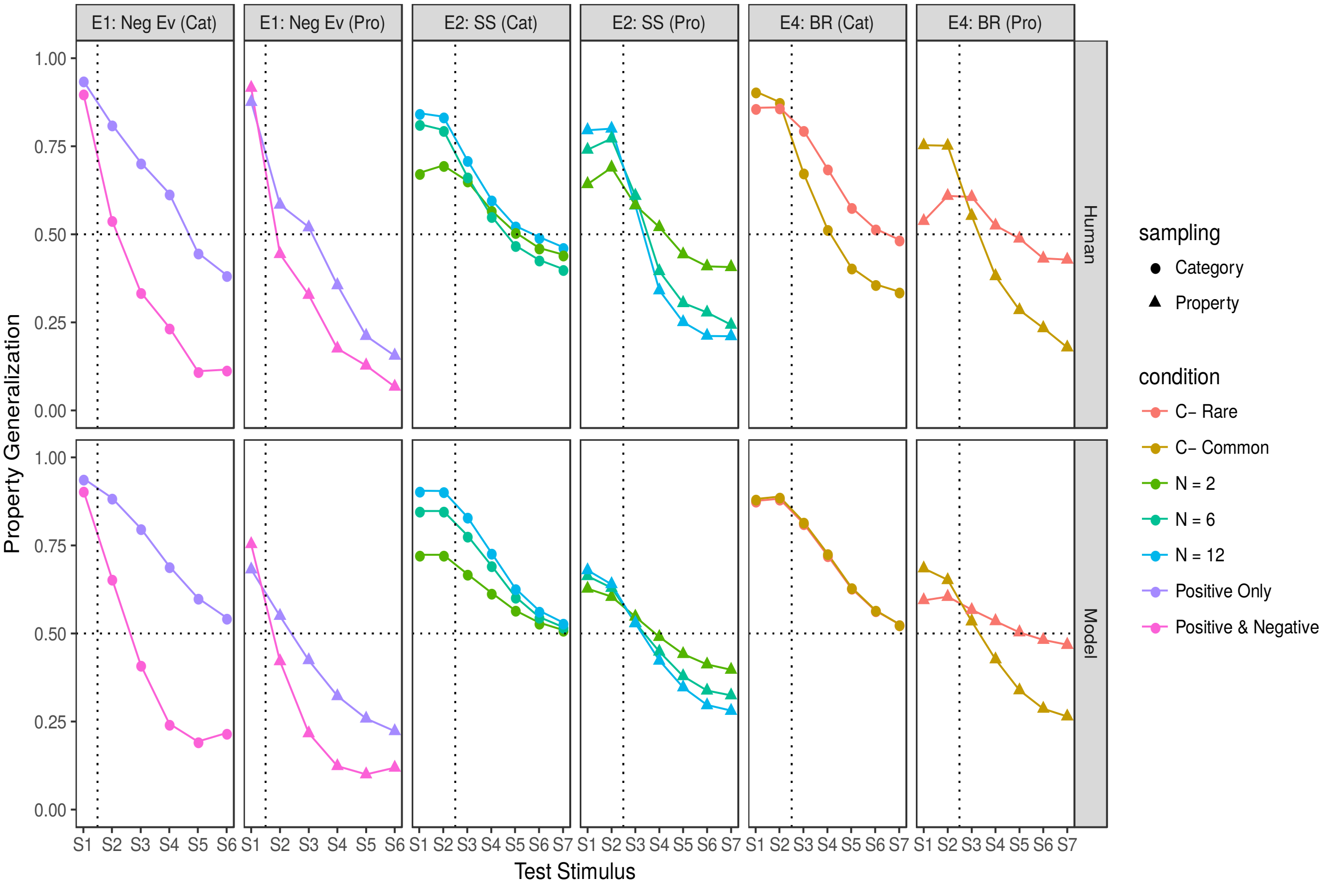


Figure 12. *Comparison between the empirical property generalization data (top row) and the predictions of the Gaussian process model (bottom row). Each panel plots the results for one of the three successful experiments (1, 2 and 4) under one of the two sampling conditions. Test stimuli for Experiment 1 correspond to the different species in decreasing order of similarity. For Experiments 2 and 4 they map to different rock sizes (with size increasing from left to right). Test stimuli to the left of the vertical dotted lines are those used in training, and those to the right are the generalization items.*

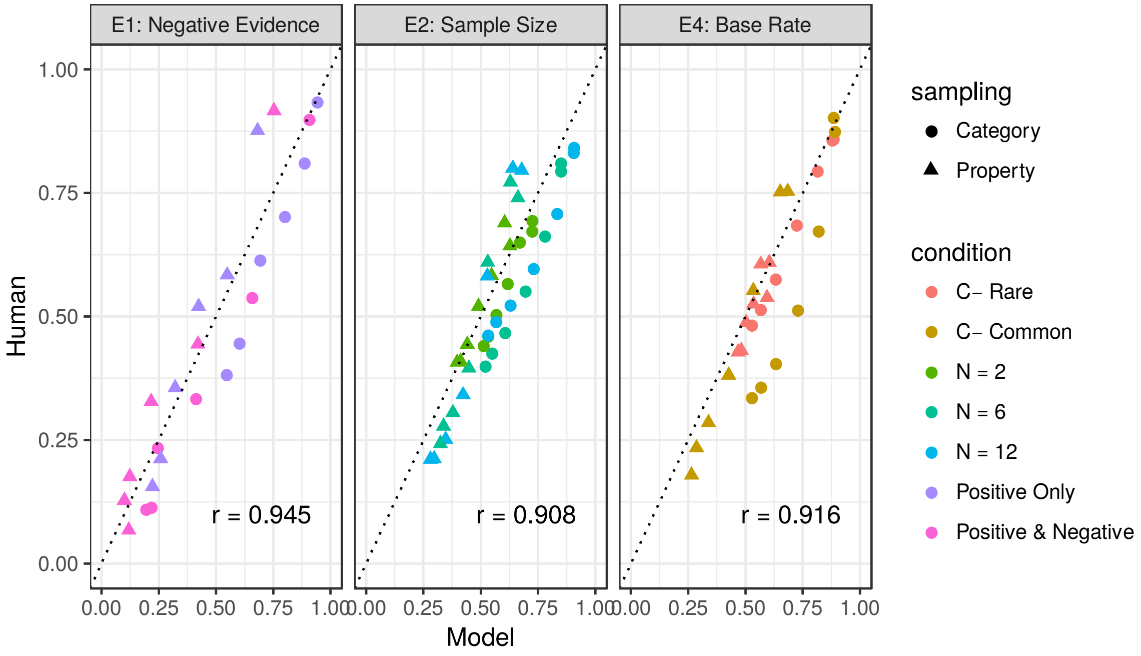


Figure 13. Scatterplots showing the comparison between the model predictions and human data for probability of property generalization (see Appendix C for the Experiment 3 plot).

7. General Discussion

These four studies examined the predictions of a new Bayesian model of property inference with censored evidence. The most central prediction from the model concerns how frames that constrain the process of sample selection affect property generalization. All four studies found evidence of a sampling frames effect, such that participants viewing identical samples that shared a novel property were more or less likely to generalize that property to other categories depending on the frame. As predicted, samples selected on the basis of a shared property generally led to more restricted property generalization than those selected on the basis of category membership. This result was robust across a range of cover stories and categorical stimuli.

We see this sample frames effect as reflecting the reasoner's sensitivity to the implications of both observed positive evidence and unobserved evidence, with frames providing alternative explanations for why certain types of data were censored. This interpretation was supported by

Experiment 1 which found that when explicit negative evidence was added to the sample, property generalization narrowed under category sampling but not property sampling. This is because the negative evidence was already strongly implied by the absence of non-target categories from the property sample. Our account was further supported by Experiments 2 and 4, which manipulated factors that affected the likelihood of observing evidence about non-target categories (sample size, category base rates). When this likelihood was reduced (due to a small sample size or low category base rate), the sample frames effect was attenuated or eliminated.

Our intuitions about the mechanisms that drive the sampling frame effect were formalized in a Bayesian account in which different censoring mechanisms (i.e. sampling frames) are implemented using likelihoods subject to different survivor functions. This model captured all of the key qualitative phenomena in our experiments. In many but not all cases, we also achieved a reasonable quantitative fit between the model and the data, even though we did not optimize most of the key parameters. Notably, our model produced a good qualitative account of inferential data involving complex multidimensional stimuli (Exp. 1) as well as unidimensional stimuli (Exp.s 2 and 4).

Most previous models of property inference (Osherson et al., 1990; Sloman, 1993), including some Bayesian accounts (Heit, 1998), have focused on the role of sample composition, with generalization governed by factors such as the typicality, size or diversity of a sample. Such models do not account for any kind of sampling effects. The current work therefore adds a new dimension to our understanding of property inference by explaining how identical samples can drive different inferences depending on perceived constraints on sample selection. In the current work these constraints were imposed by deliberate sampling strategies (e.g., reducing sampling time/complexity by only inspecting instances from a single category) or by environmental limitations (as in the size-based sampling of Sodor rocks).

This work complements recent studies that highlight the impact of social and pragmatic constraints on the sampling process in property inference (Hayes et al., in press; Ransom et al., 2016; Shafto et al., 2014; Voorspoels et al. 2015). In everyday inference the samples of evidence that we observe are often likely to be subject to both types of constraints; frame-like limitations that systematically exclude certain types of data as well as data selected with a particular social goal (e.g., to teach or to mislead). A complete understanding of property inference requires the development of detailed models of both types of data censoring processes. While such models have been developed for inferences based on socially motivated sampling (Ransom et al., 2016; Voorspoels et al. 2015), to our knowledge our model is the first to address how property inferences change as a consequence of frame constraints on evidence samples.

7.1 Comparison with other approaches

In some respects, our approach has some similarities to previous approaches that conceive of property inference as a process of constructing and assessing rival hypotheses about property extension (e.g., McDonald, Samuels, & Rispoli, 1996; Medin, Coley, Storms, & Hayes, 2003). Relevance theory (Medin et al., 2003), assumes that reasoners examine evidence samples (e.g., premise categories in verbal inductive arguments that share novel property) and construct hypotheses about what that property might be. The strongest hypothesis about property extension is based on the most salient features shared by sample instances. One crucial difference between Relevance theory and the current approach, is that in the former, hypotheses about property extension are generated by a comparison of instances *within* a given evidence sample, whereas our approach emphasizes the crucial role of external constraints (i.e., sampling frames) on the sampling process. Another crucial difference is that Relevance theory is silent about whether people consider implied (but unobserved) negative evidence when formulating inferences, whereas such evidence plays a key role in explaining the patterns of inference observed in our studies.

Throughout the paper, one key theme – though by no means the only one – is that because the sampling frame $S(d)$ constrains which potential observations d are censored and which are admissible, it plays a critical role in determining *when* "absence of evidence" should be construed as "evidence of absence". Our theoretical perspective on this question – sometimes referred to as the problem of implicit negative evidence – is explicitly Bayesian in nature, and as such it is worth considering the connection between our approach and other Bayesian perspectives on the problem. The Bayesian generalization framework (Tenenbaum & Griffiths, 2001) handles this problem by invoking the concept of *strong sampling*, in which observations are chosen explicitly from a target category. Previous work in the property induction literature (Hayes et al., in press; Ransom et al., 2016) suggests that people are sensitive to these manipulations. Although in this paper we have departed from the formalism that underpins these models (i.e., unlike Shepard, 1987, and Tenenbaum & Griffiths, 2001, we treat property inference as a function learning problem rather than inferring a consequential set of stimuli), the *censoring* mechanism can be seen as a more general version of the concept of strong sampling, one that applies across a wide variety of contexts (e.g., base rate manipulations, property versus category sampling).

Another approach to the problem arises when accounting for how people reason about verbally specified "arguments from ignorance" (Hahn, Oaksford, & Bayindir, 2005; Oaksford & Hahn, 2004; Hahn & Oaksford, 2007). An example of such an argument assumes that "If Drug A has toxic side-effects these will show up in legitimate tests". Participants are then given varying amounts of negative evidence (e.g., 1 vs. 50 tests found no side-effects) and asked to infer whether this supports the conclusion that "Drug X does not have toxic side-effects". Although such arguments are deductively invalid (equivalent to a *denial of the antecedent*), they can have inductive strength. For instance, Hahn and Oaksford found that people were more convinced by large amounts of negative evidence than small amounts. Judgments were also influenced by the strength of prior beliefs in the conclusion and beliefs about test sensitivity (i.e. the conditional probability that a side-effect would be observed if present). Hahn and Oaksford (2007) showed that these data, together with ratings of other “arguments from ignorance” that did not involve negative evidence, could be accounted for within a Bayesian framework that accommodates the idea of *epistemic closure*. In this context, epistemic closure captures an idea that is similar in spirit to the way we conceive of sampling frames – it captures the idea of completeness of the evidentiary source. If a system is epistemically closed (e.g., Google’s index of web pages) then it is expected to contain all relevant information and the absence of evidence within that system *is* informative. A system that is not closed (e.g., my browser bookmarks) is missing key information and absence of evidence cannot be construed as evidence of absence.

From the perspective advanced here, these two situations constitute different sampling frames that impose different censoring rules. Searching webpages through Google does not filter out many cases, whereas searching through one’s own bookmarks imposes very strong constraints. Broadly speaking, our approach is largely consistent with Hahn and Oaksford (2007), but is coupled with a more precise model of the inductive reasoning problem that applies in property induction tasks (i.e., Gaussian process prior over functions), and is naturally extensible to scenarios (e.g., base rate manipulations) that are not as easy to describe using an “epistemic closure” framing.

Another Bayesian perspective on the implicit negative evidence problem was proposed by Hsu et al. (2016), who present a Bayesian account of a “minesweeper” style game, in which participants had to judge whether a given area of land had been cleared of mines after inspecting a sample of locations within that area. Although the formalism they present is specific to the minesweeper task, the core ideas have a lot in common with the base rate manipulations that we used in Experiments 3 and 4, particularly Experiment 4 insofar as they also made the base rate information visually salient in the task. Like us, they find that people's inferences about implicit negative evidence are sensitive to this base rate information, but do not consider how this interacts with sampling frame. In that sense our work can be viewed as an extension and generalization of the approach.

The issue of people's sensitivity to implied negative evidence has also been examined previously in the context of inferring grammar learning, where a key question is how we learn to discriminate between grammatical and ungrammatical sequences given that the latter are rarely observed (Bowerman, 1988). Part of the answer is that language learners are sensitive to cases where an ungrammatical sequence could have been produced but was not. In these cases, the absence of the sequence implies that it is not part of the category of acceptable grammars (Hsu & Griffiths, 2009; Perfors, Tenenbaum, & Wonnacott, 2010). Our work extends these ideas by suggesting that people are sensitive to implied negative evidence when inferring how far a property generalizes from an observed sample.

A final comment on the relationship between our approaches and other Bayesian perspectives is worth adding, namely the extent to which we view our account as normative. To varying degrees, Hahn and Oaksford (2007), Tenenbaum and Griffiths (2001) and Hsu et al. (2016) suggest that normative claims are licensed from their models, insofar as they describe the model as "rational" accounts of human reasoning in different tasks. Given recent discussions on the normative status of Bayesian accounts (e.g., Bowers & Davis 2012; Griffiths, Chater, Norris, & Pouget, 2012; Tauber et al., 2017) we feel it is worth stating our own position regarding our model explicitly. Our account does satisfy a number of desirable properties of a rational reasoner (e.g., coherence), as do all Bayesian models, and can be construed as a sensible solution to the inductive inference problems we presented in our experiments. However, we stop short of making prescriptive claims about the rationality of the model and of human performance in the task. Our experiments were not designed to test whether the choices of priors and likelihoods are a good match to the kinds of problems that people face in the real world. As such, our goal was to develop a *descriptive* psychological theory about the kinds of hypotheses about property extension that people generate based on positive evidence and implied negative evidence, and how this inference process is influenced by one's understanding of mechanisms that censor some types of observations (cf. Tauber et al., 2017). How this model might translate to more complicated real world environments – and what it might say about how well people reason in them – is an open question, one to which we now turn.

7.2 Reasoning with censored evidence in more complex environments

The current work shows that the inferences we draw are sensitive to a number of factors - frames, sample size and base rates - that impact the likelihood of whether particular types of evidence will be observed during sampling. This work and most particularly, the results concerning base rates, seems to challenge strong assertions that humans rarely consider sampling constraints when drawing probabilistic inferences (e.g., Fiedler, 2012; Juslin et al., 2007; Kahneman, 2011). We see our results as an existence proof that people can combine information about base rates and sample frames to draw conclusions that are consistent with Bayesian principles. In the property induction task, our reasoners behaved more like experienced rather than "naïve" statisticians. In this respect, our conclusions are similar to those of studies examining how people make decisions based on selective feedback about decision outcomes (Elwin, Juslin, Olsson, & Enkvist, 2007; Henrik, Elwin, & Juslin, 2010). For example, people may have to learn which of a number of possible companies is most likely to yield a profit on investments. However, feedback about decision outcomes (i.e. whether a particular company made a profit or a loss) is only available for positive investment decisions– if the learner decides not to invest in the company the profit outcome is unknown. Elwin et al. (2007) found that in such situations learners use their task knowledge to *infer* the most likely outcome, resulting in decisions that provide levels of reward (investment returns) similar to those obtained when complete outcome feedback is available.

Nevertheless, some caveats on our conclusions are in order. First, we only found evidence of base rate modulation of sampling frames effects when the disparity in base rates between different categories was extreme and when this disparity was presented in a visually salient format. Second, some of the previously reported cases of base rate neglect or misuse have involved what are arguably more complex sampling scenarios than those presented in our experiments. In particular, in many cases where people have failed to factor base rates into their inferences (e.g., Fiedler, Brinkmann, Betsch, & Wild, 2000), the observed sample data have been probabilistic (e.g., only a certain percentage of the observed sample have a relevant property such as having a positive mammogram result). This contrasts with the current studies where all members of the sampled category had the property of interest. Undoubtedly, such deterministic evidence simplifies the inference process. An important goal for future work therefore is to extend the current work to cases where there is a probabilistic relationship between category membership and presence of the target property. Pilot work on this issue (Hayes et al., 2017) suggests that sampling frames effects analogous to those reported here are still observed when probabilistic evidence is used, albeit with smaller effect sizes.

The sampling frames used in the current experiments involved relatively straightforward mechanisms likely to be understood by most participants. Outside the laboratory however, people may not always understand the implications of a frame or data censoring mechanism or even be aware that censoring has taken place (e.g., Feiler, Tong, & Larrick, 2012; Hogarth et al., 2015). Returning to our management example, it may well be the case that managers often neglect the fact that previous decisions about who will be employed can produce a highly selected sample and that information about other (previously unsuccessful) applicants could be useful for predicting future job success

We also acknowledge that the sampling frames effects that were our current focus are just one of a number of different types of data censoring mechanisms that are likely to be encountered in everyday reasoning. Other forms of data censoring include truncation, where sampling methods preclude observation of instances with quantitative features that are below or above certain criteria (e.g., people below a minimum income level, people above a maximum age-threshold). The cognitive challenge in such cases is somewhat different to that studied in the current work; it involves estimating population parameters such as central tendency and variance from a truncated sample (see Feiler et al., 2013, and Koehler & Mercer, 2009 for examples). As noted earlier, we see our general framework Bayesian as potentially extensible to such tasks, but the development of detailed models of inferences based on each type of data censoring is an important task for future work.

8. Conclusion

The sampling frames effect that was focus of the current experiments is a good example of the more general problem of drawing inferences from censored data. It is rare indeed for us to have all of the available evidence at our fingertips when called upon to make everyday inferences. Instead, most of our inferences are based on data that has been subject to some form of censoring process, driven by physical and resource constraints (e.g., limited time or capacity for data accumulation) or social intentions (where data is selected by others to lead us towards a particular conclusion). The same is often true in more formal decision-making domains like finance, medicine and criminal law.

The current work extends our understanding of property inference by providing a general theoretical framework for inferences with censored evidence that allows us to predict when and how such inferences will change in different censoring environments. Our results show that when reasoners are clear about the sampling constraints that lead relevant evidence to be excluded from a sample, they do factor the missing evidence into their property inferences, in a manner consistent with our Bayesian model.

Acknowledgements

BKH and DJN devised the project, planned all experiments and carried out the modelling. SB and SF helped to design and run Experiments 1 and 4. This research was supported by Australian Research Council Discovery Grant to BKH. We would like to acknowledge the assistance of Natali Dilevski, Jeremy Ngo, Peggy Wei and Harrison Kusevskis-Hayes in data collection and coding of the experiments, and preparation of the manuscript. We thank Charles Kemp and the members of the UNSW Cognition group for useful discussions and suggestions about this research.

References

Bovens, L., & Hartmann, S. (2003). Solving the riddle of coherence. *Mind, 112* (448), 601-633.

Bowerman, M. (1988). The ‘no negative evidence’ problem: How do children avoid constructing an overly general grammar? In J. Hawkins (Ed.), *Explaining language universals* (pp. 73–101). Oxford: Blackwell.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*(3), 389-414

Chun, W. Y., & Kruglanski, A. W. (2006). The role of task demands and processing resources in the use of base-rate and individuating information. *Journal of Personality and Social Psychology, 91*(2), 205-217. http://dx.doi.org/10.1037/0022-3514.91.2.205

Elwin, E., Juslin, P., Olsson, H., & Enkvist, T. (2007). Constructivist coding: Learning from selective feedback. *Psychological Science, 18* (2), 105-110.

Feeney, A. (2017). Forty years of progress on category-based inductive reasoning. In L. J. Ball & V. A. Thompson (Eds.) *International Handbook of Thinking and Reasoning,* (pp. 189-207). Routledge.

Feiler, D., Tong, J., & Larrick, R. (2013). Biased judgment in censored environments. *Management Science, 59*, 573-591. http://dx.doi.org/10.1287/mnsc.1120.1612

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. *The Psychology of Learning and Motivation, Vol 57* (pp. 1-55). http://dx.doi.org/10.1016/B978-0-12-394293-7.00001-7

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General, 129*(3), 399-418. http://dx.doi.org/10.1037/0096-3445.129.3.399

Griffiths, T. L., Lucas, C., Williams, J., & Kalish, M. L. (2009). Modeling human function learning with Gaussian processes. In D. Koller, D. Schuurmans, Y. Bengio & L. Bottou (Eds) *Advances in Neural Information Processing Systems 21* (pp. 553-560). Curran Associates, Inc.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin, 138*(3), 415-422

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation. *Psychological Review*, *114*, 704-732. http://dx.doi.org/10.1037/0033-295X.114.3.704

Hahn, U., Oaksford, M., & Bayindir, H. (2005). How convinced should we be by negative evidence? In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 887–892). Mahwah, NJ: Erlbaum.

Hawkins, G., Hayes, B., Donkin, C., Newell, B. Pasqualino, M., & Newell, B. (2015). A Bayesian latent mixture model analysis shows that informative samples reduce base rate neglect. *Decision, 24*, 306-318. http://dx.doi.org/10.1037/dec0000024

Hayes, B. K., & Heit, E. (2018). Inductive Reasoning 2.0. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9* (3), 1-13. e1459, http://dx.doi.org/10.1002/wcs.1459

Hayes, B. K., Banner, S., & Navarro, D. J. (2017). Sampling frames, Bayesian inference and inductive reasoning. In G. Gunzelmann, A. Howes, T. Tenbrink, E. Davelaar. (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society

Hayes, B., Navarro, D., Stephens, R., Ransom, K., & Dilevski, N. (in press). The diversity effect in inductive reasoning depends on strong sampling. *Psychonomic Bulletin & Review*

Hayes, B. K., Hawkins, G. E., & Newell, B. R. (2016). Consider the alternative: The effects of causal knowledge on representing and using alternative hypotheses in judgments under uncertainty.

Journal of Experimental Psychology: Learning, Memory and Cognition, 42, 723-739. http://dx.doi.org/10.1037/xlm0000205

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford, N. Chater (Eds.), *Rational Models of Cognition*, (pp. 248-274). Oxford: Oxford University Press.

Heit, E. (2007). Models of inductive reasoning. In R. Sun (Ed). *Cambridge Handbook of Computational Psychology* (pp. 332-338).

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. (under review). Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Manuscript submitted for publication*

Henriksson, M. P., Elwin, E., & Juslin, P. (2010). What is coded into memory in the absence of outcome feedback? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36* (1), 1-16.

Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: Kind experience vs. non-transparent description. *Journal of Experimental Psychology: General*, *140*, 434–463. http://dx.doi.org/10.1037/a0023265

Hogarth, R., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science, 24*, 379-385. http://dx.doi.org/10.1177/0963721415591878

Hsu, A., & Griffiths, T. (2009). Differential use of implicit negative evidence in generative and discriminative language learning. *Neural Information Processing Systems, 22*.

Hsu, A., Horng, A., Griffiths, T., & Chater, N. (2017). When absence of evidence is evidence of absence: Rational inferences from absent data. *Cognitive Science*, *41*, 1155-1167. http://dx.doi.org/10.1111/cogs.12356

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences, 13(9)*, 381-388

Jessen, R. J. (1978). *Statistical survey techniques.* New York, NY: Wiley.

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114,* 678-703. http://dx.doi.org/10.1037/0033-295X.114.3.678

Kahneman, D. (2011). *Thinking fast and slow.* New York, NY: Farrar, Strauss & Giroux.

Kary, A., Newell, B. R., & Hayes, B K. (2018). What makes for compelling science? Evidential diversity in the evaluation of scientific arguments. *Global Environmental Change*, 49, 186-196.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90* (430), 773-395.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19*, 1-53.

Koehler J. J., & Mercer, M. (2009). Selection neglect in mutual fund advertisements. *Management Science 55*, 1107–1121.

Lawson, C. A., & Kalish, C. W. (2009). Sample selection and inductive generalization. *Memory & Cognition, 37*(5), 596-607. http://dx.doi.org/10.3758/MC.37.5.596

Lee, J. C., Lovibond, P. F., Hayes, B. K. & Navarro, D. J. (in press). Negative evidence and inductive reasoning in generalization of associative learning. *Journal of Experimental Psychology: General.*

Little, R., & Rubin, D. (2014). *Statistical analysis with missing data.* New York, NY: Wiley

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review, 22*(5), 1193-1215.

McDonald, J., Samuels, M. & Rispoli, J. (1996). A hypothesis-assessment model of categorical argument strength. *Cognition*, *59*, 199-217.

Meder, B. & Gigerenzer, G. (2014). Statistical Thinking: No One Left Behind. In E. J. Chernoff, B. Sriraman (Eds.), *Probabilistic Thinking: Presenting plural perspectives*. (pp. 127-148). Amsterdam, The Netherlands: Springer

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review, 10*(3), 517–532.

Navarro, D. J. (2018). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*. 1-7. https://doi.org/10.1007/s42113-018-0019-z

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187-223. http://dx.doi.org/10.1111/j.1551-6709.2011.01212.x

Navarro, D. J., Pitt, M. A. & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data, *Cognitive Psychology*, *49*, 47-84

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018). A formal psychological model of classification applied to natural-science category learning. *Current Directions in Psychological Science*, *27*(2), 129-135. http://dx.doi.org/10.1177/0963721417740954

Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, *58*, 75–85.

Oaksford, M., & Hahn, U. (2013). Why are we convinced by the ad hominem argument? Bayesian source reliability and pragma-dialectical discussion rules. In F. Zenker (Ed.), *Bayesian argumentation* (pp. 39-58). Springer, Dordrecht.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*(2), 185.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language, 37,* 607–642.

Pitt, M.A., Kim, K., Navarro, D. J. & Myung, J. I. (2006). Global model analysis by parameter space partitioning, *Psychological Review, 113*, 57-83.

Ransom, K., Perfors, A., & Navarro, D. J. (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, *40*(7), 1775-1796.

Rasmussen, C. E., & Williams, C. K. I., (2006). *Gaussian processes for machine learning*. Retrieved from ProQuest Ebook Central database.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R. & Wagenmakers, E. J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*, 304-321.

Sanborn, A., Griffiths, T, & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144-1167. http://dx.doi.org/10.1037/met0000057

Sanjana, N. E., & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. In *Advances in Neural Information Processing Systems* (pp. 59-66).

Schulz, E., Speekenbrink, M. & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1-16

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89. http://dx.doi.org/10.1016/j.cogpsych.2013.12.004

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25*, 213–280.

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review, 121*, 526-558.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review, 124*, 410-441

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

Wagenmakers, E. J., Ratcliff, R., Gomez, P. & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*, 28-50

Welsh, M. B. & Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes, 119,* 1-14. http://dx.doi.org/10.1016/j.obhdp.2012.04.001

Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative experience? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology, 81, 1-25.* http://dx.doi.org/10.1016/j.cogpsych.2015.07.001

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114,* 245-275. http://dx.doi.org/10.1037/0033-295X.114.2.245

Footnotes

1 Some readers may be curious about how people respond in our induction task when given no information about the sampling frame or when told that sample instances were selected randomly. We ran a pilot study using the experimental procedure described in Experiment 1 but with a cover story describing random selection of the sample instances. The pattern of generalization at test was similar to that in the category sampling – positive evidence only condition in Figure 2. At this point however, we are reluctant to draw strong conclusions from this. In debriefing it was clear that many participants no longer believed the random cover story after they observed a sample composed only of small birds.

2 The decision to adopt a within-subject design was motivated by a concern that participants might treat the sample size in a "toy" experiment in a somewhat metaphorical sense: being shown a cartoon picture of 3 rocks may not necessarily suggest to people that they should treat it as equivalent to N=3, but visually observing an *increase* in the sample size would likely prompt people to notice that the increase from 3 to 6 does correspond to an increase in evidence. Indeed, in one version of the task we ran that used a between-subjects design to manipulate sample size (Study S1 in Appendix A), we found no effect of sample size at all, though that study did replicate the sampling frames effect.

3 A rigorous comparison between the Gaussian process function learning model and a model assuming consequential regions is beyond the scope of this paper and would be ill-advised given the post-hoc nature of any such comparison using our current data sets. It is possible to apply a variation of the consequential region models to the data from Experiment 1 where there were relatively clear-cut boundaries between the various animal categories presented at test. This application does produce roughly the correct qualitative effects observed in that study (see Hayes et al., 2017). This is somewhat reassuring, insofar as it suggests that the pattern of predicted results is likely to be quite similar for any reasonable Bayesian model that incorporates an appropriate representation of the sampling mechanism. That said, we see the function learning approach as a far better alternative for cases like those in Experiments 2-4 where the boundaries between test categories are much harder to define.

Appendix A – Complete list of unreported experiments

In the interests of transparent science, and to fully disclose the contents of the "file drawer", we briefly outline the five additional experiments we have run on this topic, what those experiments found, and the reason for not including them in the main paper. All data sets, including these, are documented and included in the associated OSF repository, https://osf.io/j4dxm/ .Note that in all five cases the core "sampling frames" effect replicated, but the effects of moderating variables were not always found.

Ambiguity experiment

Experiment A1: Possible effect of ambiguous evidence. N=80 undergraduate students participated in a version of the task where 80% of the observations were P+ and 20% were ambiguous (P unknown). The frames effect replicated. Superficially, ambiguity appeared to diminish the effect, but the statistical evidence was unclear. This experiment was omitted because it relates to a somewhat different question, and moreover has previously been documented as Experiment 2 in Hayes et al. (2017).

Base rate experiments

Experiment B1: Small effect of base rate, version 1: N=286 undergraduate students participated in a version of the base rate experiment using the birds stimuli. The frames effect replicated. The predicted moderating effect of base rates appeared, but not in a robust way (i.e., the Bayes factor was modest, BF_{10} = 2.98, and the strength of evidence depended on the choice of analysis – it is somewhat larger if the analysis is based on the specific contrast we predicted rather than a general purpose ANOVA interaction, but still not entirely compelling).

Experiment B2: Minor variation of Experiment 3 in the main text (actually run before Experiment 3): N= 398 MTurk workers. The only difference in design was that there was an additional test point smaller than any observed rocks. The frames effect replicated. Consistent with Experiment 2 but not B1 there was no moderating effect of base rate.

Experiment B3: Attempted reconciliation of B1, B2 and Experiment 3 in the main. Participants were N=312 undergraduate students, who completed both the "birds" task and "robots" task in Experiment 2. The frames effect replicated in both tasks. Consistent with Experiments 2 and B2 but not B1, the moderating effect of frames was not present.

Collectively, Experiments B1-B3 were omitted because they are in minor variations of Experiment 2 in the main text, and the results are essentially replications. As noted in the main text, the base rate effect from Experiment 3 appears to be dependent on the fact that the base rate is made "sufficiently salient" using the unambiguous display shown in Figure 10.

Sample size experiment

Experiment S1: Null effect of sample size, as a between-subjects manipulation: N = 465 MTurk workers participated in a version of the "birds" task where a between-subjects manipulation of sample size was used (samples composed of 3, 8 or 20 items). The frames effect replicated. The moderating effect of sample size was not present. Unlike Experiment 2, a robust frame effect was observed for all sample sizes. This experiment was omitted for brevity, but as noted in the main text there are limits to the generalizability of Experiment 2.

Comment on possible file drawer effects

The effects reported in the paper appear to have different levels of robustness. The core “sampling frames” effect was tested in all 9 experiments and was detected on all 9 occasions. The interactions between frame and other variables may be less robust than the core effect. To summarize:

- Interaction with negative evidence: tested once (Exp. 1), detected once
- Interaction with ambiguity: tested once (A1), detected once but with small effect
- Interaction with sample size: tested once (Exp. 2) within-subjects and detected as a large effect; once between-subjects and not detected
- Interaction with base rate: tested five times (Exp. 3-4, B1-B3), one strong effect with a very salient manipulation (Exp. 4), no consistent pattern of evidence with less salient manipulations (other experiments)

Appendix B - Robustness and model complexity

As mentioned in the main text, parameter estimation was done by hand rather than via an explicit optimization procedure. Visual inspection of the curves during the parameter tuning process suggested that the theoretically important qualitative trends in the data were almost always reproduced by the model regardless of the parameter values chosen. To substantiate this intuition a little more quantitatively, we undertook a model evaluation procedure – the logic for which is discussed by Navarro (2018) – that is a hybrid of the parametric bootstrap (see Wagenmakers, Ratcliff, Gomez & Iverson, 2004), landscaping (Navarro, et al., 2004), and parameter space partitioning (Pitt, et al., 2006). We sampled 1000 parameter values randomly from fairly diffuse, quasi-informed prior distributions centered on the best fitting parameter values, and for each parameter set determined whether the theoretically-relevant effect is reproduced by the model. For σ, τ, ρ and α, we sampled from an exponential distribution with mean centered on the parameters reported in the text; μ was drawn from a standard normal; and for the base rate manipulations the value of α for the rare and common categories were multiplied (or divided) by scaling factors sampled uniformly from 1 to 100. We evaluated 14 qualitative contrasts. Unless otherwise stated, the expected base rate of success by chance is 50%.

- Experiment 1: Overall generalization decreases when explicit negative evidence is added – found in 998 of 1000 cases
- Experiment 1: Overall generalization is lower in property than category sampling – found in 967 of 1000 cases
- Experiment 1: The difference between category and property is attenuated when negative evidence is added – found in 994 of 1000 cases.
- Experiment 2: Under category sampling, sample size increase shifts generalization upwards overall – found in 989 of 1000 cases (against chance base rate: 16.7%)

- Experiment 2: Under category sampling, the upward shift is larger for target categories than dissimilar categories – found in 977 of 1000 cases (against chance base rate: 16.7%)
- Experiment 2: Under property sampling, sample size increase shifts generalization upwards for target categories – found in 593 of 1000 cases (against chance base rate: 16.7%)
- Experiment 2: Under property sampling, sample size increase shifts generalization downwards for dissimilar categories – – found in 759 of 1000 cases (against chance base rate: 16.7%)
- Experiment 2: Overall generalization is lower in property than category sampling – found 997 of 1000 cases
- Experiment 4: Overall generalization is lower in property than category sampling – found 1000 of 1000 cases
- Experiment 4: Under property sampling, shifting base rate of C- from rare to common increases endorsement of C+ – found 660 of 1000 cases
- Experiment 4: Under property sampling, shifting base rate of C- from rare to common decreases endorsement of C- – found 930 of 1000 cases
- Experiment 4: Under category sampling, shifting base rate of C- from rare to common increases endorsement of C+ – found 496 of 1000 cases. This, as well as the next prediction in the list, represents the null prediction of the model about the effects of the base rate manipulation on category sampling. We acknowledge that this prediction was not supported by the data (see Section 6.2)
- Experiment 4: Under category sampling, shifting base rate of C- from rare to common decreases endorsement of C- – found in 477 of 1000 cases
- Experiment 4: The magnitude of the base rate by sampling crossover effect is smaller in category sampling than property sampling – found in 940 of 1000 cases

These results make clear that the many successes of the model (almost all of the effects listed above) and the occasional failures (e.g., inability to produce an attenuated rather than a null base rate effect in category sampling) are *generic* predictions of the model. Indeed, in most cases it is almost impossible for the model to make any prediction other than what we found in the data. In short, it seems unlikely that the strong performance of the model is an artifact of model complexity.

Appendix C – Modeling Experiment 3

The Bayesian model predictions were identical for Experiments 3 and 4. Hence we compared these predictions against the observed data for Experiment 3 (Figure C1). The overall model fit to the data is still reasonable (*r* = 0.84), albeit lower than was found for the other experiments. Consistent with our earlier discussion of these results, the most notable model misfits involve underestimation of property generalization in category and property sampling when C- was rare.

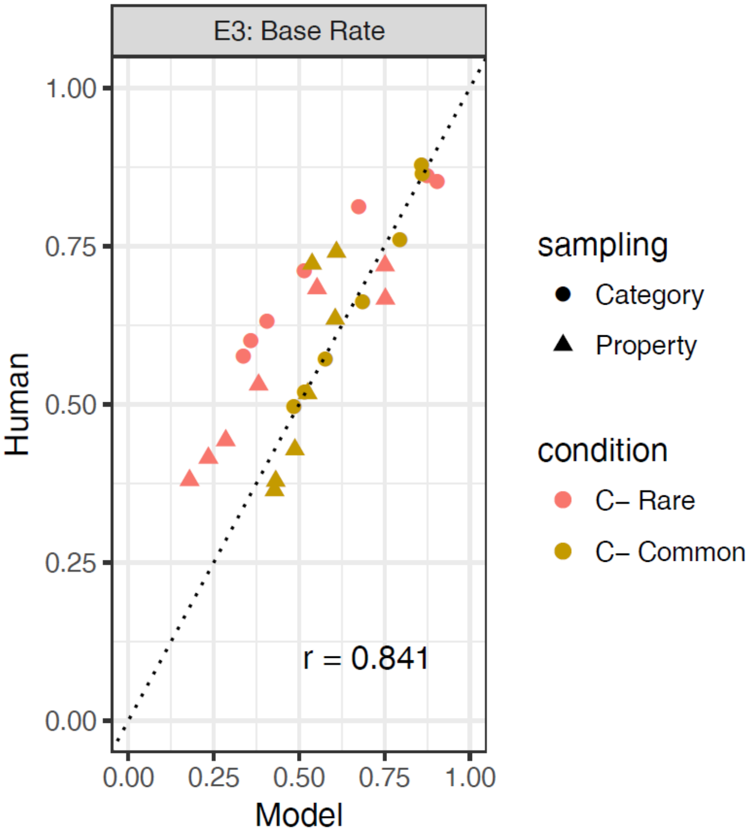


Figure C1. Scatterplot showing the comparison between the model predictions and human data for probability of property generalization in Experiment 3.