Ransom, K. J., Perfors, A., Hayes, B. K., & Connor Desai, S. (in press). What do our sampling assumptions affect: how we encode data or how we reason from it?. *Journal of Experimental Psychology: Learning, Memory and Cognition.* https://doi.org/10.31234/osf.io/yqtd9

What do our sampling assumptions affect: how we encode data or how we reason from it?

Keith J. Ransom[1], Andrew Perfors[1], Brett K. Hayes[2], and Saoirse Connor Desai[2]
[1]School of Psychological Sciences, University of Melbourne
[2]School of Psychology, University of New South Wales

Abstract

In describing how people generalize from observed samples of data to novel cases, theories of inductive inference have emphasized the learner's reliance on the contents of the sample. More recently, a growing body of literature suggests that different assumptions about how a data sample was generated can lead the learner to draw qualitatively distinct inferences on the basis of the same observations. Yet relatively little is known about how and when these two sources of evidence are combined. Do sampling assumptions affect how the sample contents are encoded, or is any influence exerted only at the point of retrieval when a decision is to be made? We report two experiments aimed at exploring this issue. By systematically varying both the sampling cover story and whether it is given *before* or *after* the training stimuli we are able to determine whether encoding or retrieval issues drive the impact of sampling assumptions. We find that the sampling cover story affects generalization when it is presented before the training stimuli, but not after, which suggests that sampling assumptions are integrated during encoding.

Keywords: Bayesian reasoning, inductive reasoning, generalization, sampling assumptions, category learning

Introduction

Theoretical accounts of inductive inference – how learners generalize from samples of evidence to novel cases – have traditionally focused on the sample content: what data are observed (see Feeney, 2018; Hayes, in press, for reviews). Imagine, for example, that your roommate is making

Keith J. Ransom https://orcid.org/0000-0001-5423-6455
Data are available from the corresponding author via request.
Correspondence concerning this article should be addressed to Keith J. Ransom, School of Psychological Sciences, University of Melbourne, Melbourne, VIC 3010. E-mail: keith.ransom@unimelb.edu.au

a playlist on their new Spotify account. Watching them find and add songs by Nirvana and Pearl Jam, might lead you to infer that they like 1990's grunge bands. Seeing them next add a track by Soundgarden might strengthen this belief. Such inferences are often explained by computing the similarity between the observed sample and generalization targets and leveraging prior knowledge of categorical relations (see, e.g., Osherson et al., 1990).

Generalizing appropriately from samples of evidence, however, should also involve a consideration of how the sample was generated in the first place. You would probably make weaker inferences about your roommate's music preferences if instead of directly choosing songs to add to their playlist, they had been listening to a random shuffle from the Spotify catalogue and adding the songs that they liked to their playlist along the way. If you learned that they were screening songs from a live stream of all-male bands, your inference might be weaker still.

These examples illustrate that the evidential weight that a learner attaches to sample observations should depend on the learner's beliefs about the data generation process – that is, on their *sampling assumptions*. This dependence has been investigated within two related lines of research. One line of work seeks to understand sampling assumptions arising in social contexts, with typical experiments contrasting inferences based on *strong* and *weak* sampling. Under strong sampling, samples are selected by an intentional agent because they are known to share some interesting property (e.g., your roommate specifically chose the grunge songs to add to their playlist because they like those songs). Under weak sampling, by contrast, sample selection is not directly connected with the property or concept that the learner is considering. Instead, samples are generated in a way that is unconstrained by, uncorrelated with, or otherwise independent of the learner's goal (e.g., if your roommate is choosing from songs being sampled by a random algorithm from a large and diverse collection). Applying different sampling assumptions to the same data can lead to qualitatively distinct patterns of inductive generalization (Hayes, Navarro, et al., 2019; Hendrickson et al., 2019; Ransom et al., 2016), shifts in epistemic trust (Mascaro & Sperber, 2009; Shafto et al., 2012), changes in pragmatic implicatures (Goodman & Frank, 2016) and the promotion of learning (Shafto et al., 2014).

In contrast to the study of intentional sampling mechanisms, a second line of research has focused on assumptions about *sampling frames*: unintentional constraints that arise from the structure of the learning environment (e.g., Hayes et al., 2017; Hayes, Banner, Forrester, et al., 2019; Lawson & Kalish, 2009). Such constraints mean that only certain types of instances are admitted to the sample, while others are systematically excluded. The crucial intuition is that the inferences we draw from a data sample will depend on our beliefs about what types of constraints have been applied to that sample. In our example, you could not draw many conclusions about whether your roommate was a particular fan of 1990's grunge if you knew that the songs were selected from a live stream that happened to be playing only all-male bands. This example illustrates what has been referred to as *category sampling*, where selection constraints mean that only members of a particular sub-category can be observed in the sample. Crucially, the absence of any sample instances other than all-male bands is entirely attributable to the frame – the data tells us nothing about your

roommate's actual music preferences. In contrast, your conclusions would be far stronger if you knew that the songs being played came from music stored locally on your roommate's device. This type of frame has been referred to as *property sampling*, and yields a stronger inference: if it was theoretically possible to observe other types of bands in the sample, their absence is informative.

The impact of sampling frames on inductive inferences is often studied by presenting identical samples to learners but varying what they are told about how the samples were generated. For example, in Hayes et al. (2017) and Hayes, Banner, Forrester, et al. (2019) all learners saw a sample of instances from a single category (small birds) with a novel property ("plaxium blood"). The explanation of sample selection was varied between groups; sample instances were selected either because only small birds were available for inspection (CATEGORY frame) or because they were the first items to have passed a screening test for the presence of the property (PROPERTY frame). The respective frames led to different patterns of inductive generalization. Those given the category frame generalized more broadly, extending the property to related but novel categories like large birds; those given the property frame showed narrower generalization, with little extension of the property beyond the observed sample.

Despite the important role of sampling assumptions in generalization, relatively little is known about the specific processes that mediate these effects. One possibility is that learners' assumptions about sampling processes affect the way that the data are learned or *encoded* in the first place, shaping the way that the evidence is represented in memory. Alternately, people with different sampling assumptions may encode evidence in the same way but use that evidence in different ways when called on to make inferences about novel cases. According to this latter possibility, sampling assumptions would be relevant only during *retrieval* when generalization judgments are required. The goal of the current work was to experimentally investigate these two alternative accounts and thus illuminate part of the process by which assumptions about data generation guide generalization.

Our approach to this issue is straightforward. As detailed in the next section, the encoding- and retrieval-based accounts lead to divergent predictions about how generalization will be affected based on *when* the sampling assumptions are made explicit (i.e., whether before or after a data sample is observed). We examine these predictions using each of the two major inductive paradigms previously used to study sampling assumptions. In Experiment 1 we manipulate beliefs about whether the data in a single-category generalization task were generated by strong or weak sampling. In Experiment 2 we manipulate the frames that determine how data is sampled in a property induction task. Both experiments support the hypothesis that sampling assumptions shape generalization through encoding rather than retrieval processes.

Formalizing sampling assumptions

Sampling assumptions in single-category generalization can be captured within a Bayesian framework (e.g., Tenenbaum & Griffiths, 2001) which presumes that the learner is deciding among a set of hypotheses $\mathcal{H}_c$ about the true extension of a category c based on previously observed exemplars of that category, $\mathbf{d}$. The posterior probability $P(h \mid \mathbf{d}, s)$ for each hypothesis h given the data

d and a sampling assumption s depends on two quantities: the prior probability of that hypothesis $P(h)$ and the likelihood $P(\mathbf{d} \mid h, s)$ of the data given that hypothesis and the sampling assumption.

Within this framework, it is the likelihood function that is crucial for capturing the effects of different sampling assumptions. As Appendix A shows in detail, different likelihoods yield different predictions about how the learner should generalize. For example, under **weak sampling**, data which are incompatible with a hypothesis have a probability of zero and data that are compatible with it have a probability given by the base rate of the exemplars involved (which is independent of the hypothesis itself). A consequence of this assumption is that the observed data provides no clue as to which of the data-compatible hypotheses is correct. As a result, observing additional weakly-sampled data should not change generalization unless it is actually incompatible with the hypotheses under consideration.

In contrast, under **strong sampling** the same data *is* informative for choosing between data-compatible hypotheses. This is because the likelihood of seeing an exemplar x consistent with a hypothesis h is inversely proportional to the size or extent of that hypothesis (denoted $|h|$). By extension, if the data **d** consist of n independently sampled observations then this size is raised to the nth power ($|h|^n$). An important implication of strong sampling is the *size principle*: smaller hypotheses will generally be more probable than broader hypotheses, for a given sample of data. This captures the intuition that more specific hypotheses offer a better "fit" to data than hypotheses which are consistent with nearly anything. For the learner it implies that with each additional observation a further "tightening" of generalization is warranted whereby smaller hypotheses consistent with the data are regarded as even *more* likely.

Although the specifics differ, a broadly similar approach has been suggested for modeling generalization based on property and category sampling in property inference (Hayes et al., 2017; Hayes, Banner, Forrester, et al., 2019). In this case, people are asked to generalize about which categories have a certain property; the sampling assumption or sampling frame s thus represents a "censoring function" which only allows certain types of data to be observed. The effects of the different sampling assumptions are again implemented via different likelihood functions. In **category sampling**, the censoring function restricts observations to a subset of the categories of interest (e.g., only small birds). The data are therefore uninformative about whether the property in question applies to the remaining categories (e.g., large birds): the fact that all observations happen to belong to a limited set of categories has no evidentiary value because the sampling frame s only allowed those types of items in the first place. As with weak sampling, because the data are uninformative no tightening of generalization with additional data points is warranted.

Under **property sampling**, the sampling frame admits observations from any category as long as they have the target property. The likelihood function in this case is a straightforward extension of strong sampling with smaller hypotheses (which suggest that fewer items have that property) being more strongly supported by the same data. As with strong sampling, we therefore would expect tighter generalization with increasing data under conditions of property sampling.

The predicted differences in the tightening of generalization under strong and weak sam-

pling assumptions have been observed across of range of single-category learning and property inference tasks (e.g., Hayes, Banner, Forrester, et al., 2019; Hendrickson et al., 2019; Navarro et al., 2012; Ransom et al., 2018; Ransom et al., 2016; Xie et al., 2020). Likewise, the predicted differences between property generalization under property and category sampling have also been demonstrated in property inference (Hayes, Banner, Forrester, et al., 2019). What remains unclear, however, is *when* the crucial likelihood functions are calculated and incorporated into the learner's inferences: at encoding, or at retrieval?

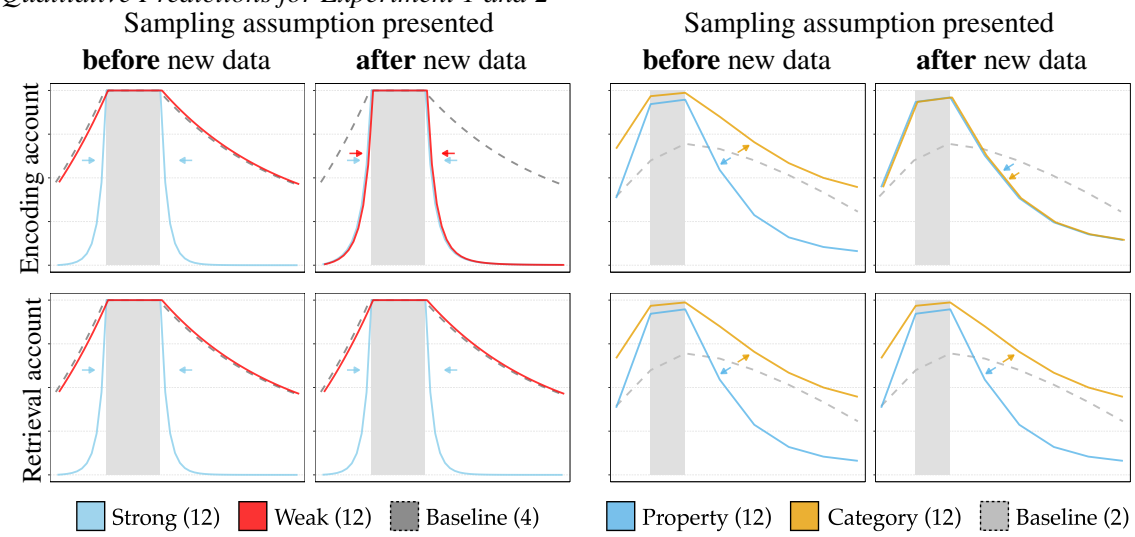
This issue has not been the focus of previous work studying inductive generalization within a Bayesian framework. In part, this is because the Bayesian approach to inductive inference is frequently characterized as a computational or "as if" model of how generalization from samples of evidence should turn out (cf. Anderson, 1991; Marr, 1982). Such approaches are agnostic about when and how specific computational operations (e.g., calculating likelihoods) occur during the course of the generalization process. In contrast, our ultimate goal is to develop a detailed, descriptive model of how learners actually approach the task of inductive generalization (cf. Tauber et al., 2017). Discovering whether sampling assumptions operate primarily at the point of encoding or retrieval is an important step towards that goal. As we explain below (and show in Figure 1), each possibility leads to a different prediction about the pattern of generalization one would expect as a function of when the sampling assumptions were made explicit.

Encoding account. One possibility is that the likelihood function is applied as each exemplar (i.e., each data point or sample instance) is encountered. This information could be represented in a number of ways. For example, one possibility is that the likelihood for each hypothesis is stored separately for each instance, and a sample of this evidence is later accessed when instances are retrieved during generalization. Alternatively, the likelihood for each hypothesis could be aggregated across instances and represented as a single set of weights. Here we do not attempt to differentiate these variants of the encoding account; regardless of these specifics, if the likelihood is calculated at encoding, then generalization judgments should be shaped by the sampling assumptions that applied at that time. This should be true even if those assumptions are revised after the data was encoded. Hence, an encoding account would predict tighter generalization under strong sampling compared to weak sampling, but *only* if the sampling assumptions were made explicit before the data were shown; otherwise, generalization should be identical. A parallel prediction applies to property generalization: generalization should be tighter for property sampling than for category sampling only if the sampling assumptions were made explicit before observing the data.

Retrieval account. An alternative possibility is that the likelihood is not calculated until generalizations about novel items are required. During generalization, likelihoods are applied to the instances of the sample that are retrieved (which may be a smaller set than the total observations if some are forgotten). According to a retrieval account, the way that data is encoded is unaffected by the learner's assumptions about the sampling process. Hence, whether the data generation process is explained before or after the data are presented (i.e. before or after encoding), the effect on subsequent generalization should be the same regardless: in either case, generalization should be

Figure 1

Qualitative Predictions for Experiment 1 and 2



(a) Experiment 1 predictions (b) Experiment 2 predictions

Note. This figure illustrates the generalization patterns predicted by the encoding and retrieval accounts as a function of whether the sampling assumption is made explicit before the data are observed or after. Each panel shows predicted generalization gradients under the two respective sampling manipulations (solid lines) and at baseline (dashed line) for Experiments 1 and 2 (number of training items for each condition appear in parentheses; the grey rectangles indicate the range of training data observed). The curves represents the probability (y-axis) that a given exemplar (x-axis) is an instance of the concept of interest. When the sampling assumption is made explicit before the data are presented, then both accounts make identical predictions that each sampling assumption will induce its own characteristic pattern of generalization. In contrast, when the sampling mechanism is made clear only after the data is observed, then the predictions of the two accounts diverge. In this case, the encoding account predicts that generalization will be unaffected by whichever sampling assumption is made explicit after encoding, being driven instead by the "default" likelihood calculated at the time of encoding (for the sake of illustration, an even mixture of the two explicit likelihoods is assumed). Under the retrieval account, generalization is governed by the likelihood calculated at the point of retrieval. Thus it makes the same predictions regardless of whether the sampling assumption was made explicit before or after the data were observed: namely that strong sampling and property sampling reduce generalization to novel items outside of the range of data observed during training.

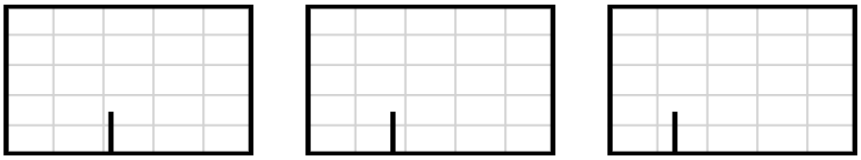
tighter under strong or property sampling compared to weak or category sampling.

Experiment 1

The goal of the present study is to investigate how and when sampling assumptions mediate generalization behaviour. Our first experiment aimed to distinguish between encoding and retrieval accounts using a single-category learning task. Participants viewed instances of a novel category and

Figure 2

Example Stimuli from Experiment 1



Note. Items varied only in the position of the short black vertical line along the bottom edge of the rectangle.

then judged whether transfer items belonged to the same category. The experiment was modeled on previous work demonstrating that sample size and sampling cover story affect people's willingness to extend category membership to novel examples (Hendrickson et al., 2019; Ransom et al., 2018). As in those experiments, one of our manipulations involved the nature of the cover story people received. Either they were told that the data was given by a helpful teacher (which corresponds to a STRONG sampling assumption and implies that generalization should be tighter) or they were given a cover story and a learning context implying that it was chosen at random (which corresponds to a WEAK sampling assumption and implies that generalization should be looser). Critically, we manipulated whether people were given the sampling information BEFORE or AFTER they saw the training stimuli. If sampling assumptions affect how the data are encoded then people should generalize differently depending on when they received this information.

Participants

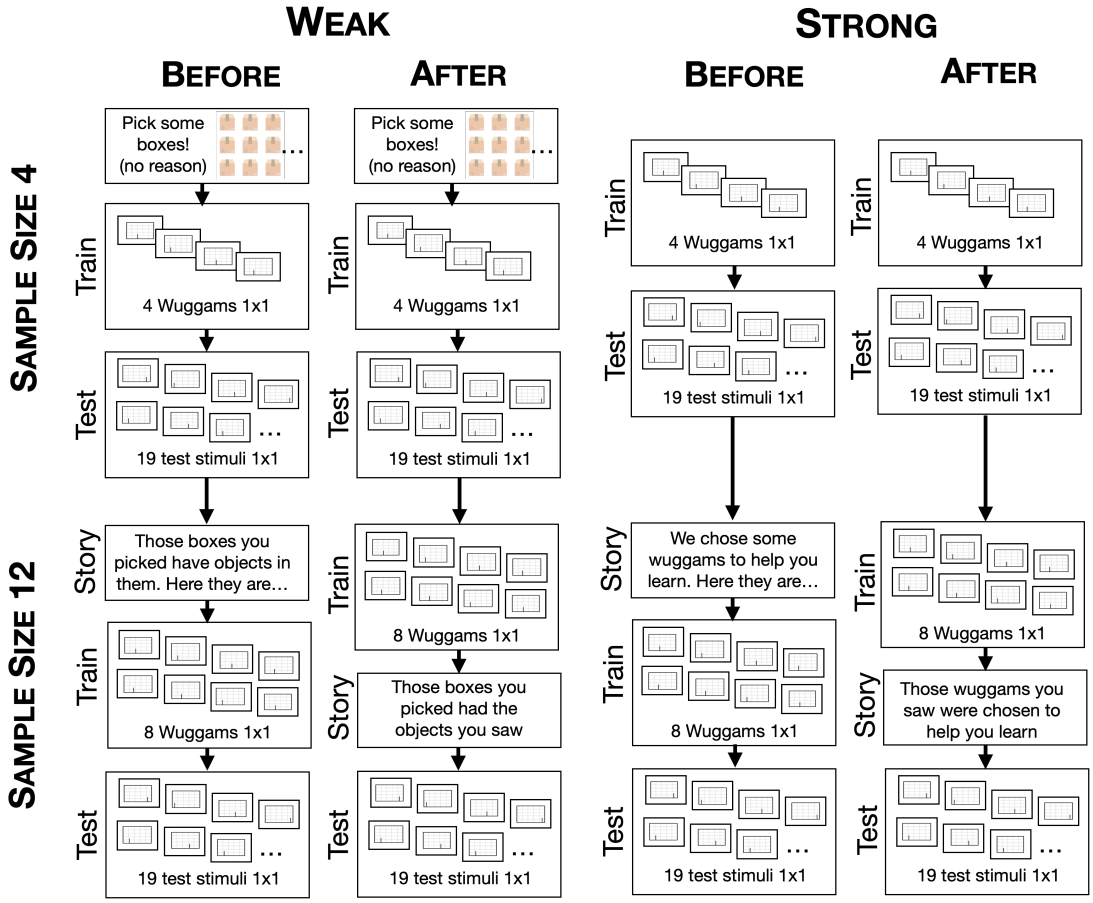
We recruited 999 people via Amazon Mechanical Turk who were each paid $1.70USD for 5-10 minutes participation. 56% were female, with age varying between 18 and 75 (median: 37 years), drawn predominately from the U.S. population (99%). All participants passed a screening for English language competency prior to participation and indicated informed consent via an online consent form. The research was approved by the School of Psychology Human Research Ethics Subcommittee of the University of Adelaide (approval code: 16/94). All procedures were performed in accordance with the institutional guidelines.

Stimuli

Stimuli were black rectangles containing a vertical black line inside, attached to the bottom edge (see Figure 2). They varied along a single dimension (the *stimulus value*): the horizontal position of the line within the rectangle. Participants were told that this was the way in which stimuli varied. Evenly spaced light grey "guide lines" were drawn within each rectangle in order to improve discriminability. There were 12 training stimuli in total, whose stimulus values ranged from 21% to 43% in increments of 2%. They were divided into two sets corresponding to the two training phases, as described below.

Figure 3

Design for Experiment 1



Note. Our 2×2×2 design varied Sample Size within-subject and Sampling Explanation and Presentation Sequence between-subjects. All participants began by seeing four individually-presented exemplars followed by a generalization task to novel stimuli. Those in the WEAK condition were then given a cover story in which the subsequent eight items were chosen at random from boxes that they themselves had previously selected. Those in the STRONG condition were told that the items were selected by a helpful teacher. In the BEFORE condition, the cover story was given before seeing the eight new items; in the AFTER, it came after. In all conditions the experiment ended with a repeat of the generalization test.

Design and procedure

As shown in Figure 3, our experiment employed a $2 \times 2 \times 2$ mixed factorial design. Two factors (Sampling Explanation and Presentation Sequence) were manipulated between-subjects while another (Sample Size) was varied within-subjects. People were thus allocated at random to one of four experimental groups. Across all groups, the experiment involved presenting people with a number of examples of a novel one-dimensional category and then observing whether they general-

ized category membership to new items based on the examples they had been shown and what they had been told about how those examples were generated.

Sample Size To facilitate a baseline against which the effect of additional exemplars could be compared, the experiment involved two rounds of testing. The first (Size 4) occurred after a training phase involving four training examples, and the second (Size 12) after seeing eight more.

Stimuli for the first training phase consisted of the two extreme examples (with values of 21% and 43%) and two others selected at random from the ten between those extremes. The eight remaining stimuli formed the second training set and were presented in random order.

Presentation Sequence This between-subjects manipulation varied when the sampling cover story was presented in relation to the second training set. People in the BEFORE condition were told the cover story (WEAK or STRONG, described below) *before* viewing the second set of training items, while people in the AFTER condition were offered the explanation only after all training items had been presented.

Sampling Explanation The other between-subjects manipulation varied the explanation given for how the data in the second training phase were generated. The initial training phase, however, was identical for all participants. No explanation was given for how the exemplars were chosen. People were told only that the purpose of the experiment was to see how people judged whether or not unfamiliar objects were in the same category as known examples. In the second training phase people were given one of two different cover stories explaining how the items were selected.

Strong. People in the STRONG condition were told:

> We have a bunch of boxes containing examples of the full variety of «Wuggams». We have chosen 8 of these boxes especially to help you learn the «Wuggam» category, bearing in mind the four training examples we showed you originally.

at which point an array of eight icons resembling open packing boxes were displayed in an adjacent panel. Participants in the BEFORE condition then viewed the eight stimuli one by one. Those in the AFTER condition saw the identical explanation (with verb tenses adjusted) only after all eight stimuli in the second training phase had been shown.

Weak. The WEAK condition was designed to dissaude people from concluding that the examples shown reflected a representative or unbiased sample of the category of interest. Instead, people were encouraged to believe that each training item was selected from a diverse (and ill-specified) set of items which also contained items that were not examples of the target category. Previous work suggests that people tend to assume strong sampling by default (particularly when no examples falling outside the category are presented); as a result, instructional manipulations alone are often insufficient to suppress a belief in strong sampling or intentional selection of data (e.g., Hayes, Navarro, et al., 2019; Ransom et al., 2016). In order to make the weak sampling cover story more believable, we therefore combined it with a learning context that reinforced the existence of a weak sampling process. We did this by presenting people in the WEAK condition with an additional phase preliminary to the first training round in which they were shown a 6×5

arrangement of packing boxes and asked to select boxes in any order (without being told why this was necessary). After selecting 11 boxes, people were told that the contents would be revealed later in the experiment. Following this, the first training phase commenced, which was identical for all participants.

At the start of the second training phase participants were informed that they would be given a chance to learn more about which items were Wuggams. Following this, participants in the AFTER condition were immediately shown the eight remaining training items without explanation. People in the BEFORE condition were told that we had many boxes containing examples from our catalogue, and that these examples included but were not limited to Wuggams (this is important because under weak sampling it is possible to sample instances from outside of the category). After this, the original array of (closed) boxes was displayed, indicating the ones that the participant had previously selected. People were then told:

> At the start of the experiment we asked you to choose some of these boxes at random. These are the boxes that you selected. We're going to open them now and show you whatever kind of item we find inside.

In order to reinforce the notion that it might have been possible to see items from categories other than Wuggams, the display was updated at this point to reveal eight open boxes and three closed ones. People were told that some of the boxes they had chosen were stuck but that we would show them the contents of the boxes that did open. Participants in the AFTER condition received exactly this cover story (with verb tenses adjusted) only *after* seeing all eight training examples.

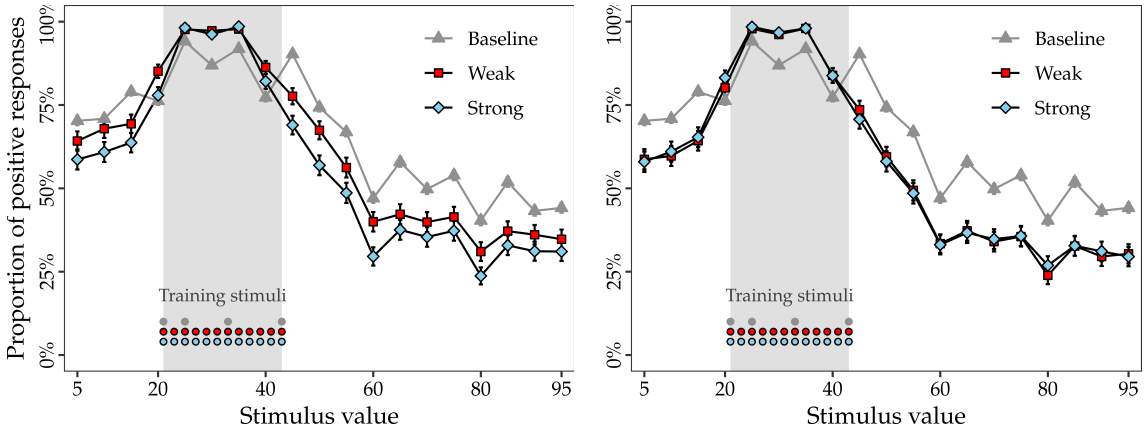
Generalization test. Immediately after both the first and second training phase, participants in all conditions performed the same generalization test. They were shown 19 stimuli one at a time in random order; this sequence was repeated four times. The stimuli consisted of 19 items with stimulus values ranging from 5% to 95% in increments of 5%. The test query was a yes or no question: "Do you think this object is in the «`Wuggam`» category?" Neither training stimuli nor the sampling explanation remained on-screen during testing, requiring people to rely on their memory when making judgements.

Results and discussion

Our analyses are organised around two key questions. Do we replicate previous findings showing that differences in sampling assumptions lead to differences in generalization? And importantly, how does any effect observed depend on the stage at which the sampling explanation was provided? In considering this second question, we attempt to distinguish between three qualitatively different patterns of effect: *Retrieval only*: if generalization is influenced only by the sampling assumption which holds at the time that the generalization target is assessed, we should expect the same patterns of generalization in both the BEFORE and AFTER conditions (with WEAK BEFORE matching WEAK AFTER and STRONG BEFORE matching STRONG AFTER); *Encoding only*: in contrast, if people's sampling assumptions take effect by changing the way that observations are

Figure 4

Category Generalization as a Function of Presentation Sequence, Sampling Explanation and Sample Size



(a) BEFORE. (b) AFTER.

Note. The figure illustrates people's performance on the one category generalization task featured in Experiment 1. The plots show the proportion of positive responses to the question: "Do you think this object is in the «Wuggam» category?" for each of the test stimuli. People's performance after seeing four examples of the target category with no sampling explanation given (grey line) is contrasted with their performance after seeing all 12 examples and being given an explanation of how the additional examples were selected (black lines). (a) When the sampling explanation was given prior to the presentation of the final 8 examples (BEFORE condition), people tightened their generalizations as more data was observed, but the extent of tightening was affected by the sampling manipulation; those people who actively sampled the additional examples at random (red squares) tightened their generalization less than those that were told that the items had been selected by a helpful teacher (blue diamonds). (b) In contrast, when the sampling explanation was given only after all training stimuli were presented (AFTER condition), the sampling manipulation had no effect, with people tightening their generalization equally in both cases.

encoded, and subsequent updating of assumptions has no effect on later generalization, then we should expect that a) there will be more tightening of generalization in the STRONG BEFORE than the WEAK BEFORE conditions, and b) equivalent patterns of generalization in the WEAK AFTER and STRONG AFTER conditions (since there should be no systematic differences regarding the "default" sampling assumption that people held when observations were encoded); *Encoding and retrieval*: a third possibility of interest is that people's generalizations are influenced by the sampling assumptions that hold at the time that the generalization target is assessed, but there is also continued influence of the assumptions adopted when data was encoded. A reduced effect of sampling explanation in the AFTER condition compared to the BEFORE condition would be consistent with this possibility.

To address these questions, we constructed the generalization gradients shown in Figure 4 which aggregate the responses of participants separately for the first and second and test phases.

Table 1

Comparison of how well three different regression models capture the responses from the second test phase of Experiment 1 (lower LOOIC indicates better fit).

Model	Model performance				
	LOOIC	SE	Contrast	$\text{LOOIC}_{\text{diff}}$	SE_{diff}
1. BASELINE	16144	250	–	–	–
2. +MANIPULATION	16147	250	LOOIC_{2-1}	3.5	2.4
3. FULL	16276	253	LOOIC_{3-2}	128.3	16.9

Note. We consider three theoretically motivated models. Each model connects a beta-binomial response to a linear combination of predictors via a logistic link function, and all models include a random intercept for each individual. The BASELINE model is intended to capture a change in generalization gradient between test phases where the degree of change varies according to the stimulus. The model thus includes the test stimulus and the corresponding response count (observed during the first test phase) as predictors, as well as an interaction between the two. The +MANIPULATION model adds predictors for sampling explanation and presentation sequence (with an interaction between the two), designed to reflect a uniform change in response threshold driven by our experimental manipulation. By including all interaction terms, the FULL model is able to model an effect of our experimental manipulation that varies in strength according to the test stimulus, in line with the predictions of the Bayesian generalization model. Comparable fits for the BASELINE and +MANIPULATION models raise the possibility that the experimental manipulation had a modest effect.

The baseline (shown in grey) captures responses from all participants having seen only the first four training items for which no sampling explanation was given. The remaining gradients (shown in black) represent responses made after participants had viewed additional training stimuli for which an explanation was provided: either that items were sampled helpfully with a view to supporting the generalization in question (red squares), or selected at random (blue diamonds). Compared to the baseline, the patterns of generalization observed across conditions are suggestive of a clear effect, where observing additional stimulus items makes people less willing to generalize a category outside of the narrow range observed. Such tightening of generalization, is consistent with learners having made some form of strong sampling assumption. The plots suggest a partial replication of previous results (notably Ransom et al., 2018), depending on *when* the sampling explanation was given. In the BEFORE condition (panel a), where people were given the sampling explanation before seeing the additional training stimuli, the STRONG explanation leads to narrower generalization than the WEAK explanation.[1] In contrast, when participants received the sampling explanation after the training stimuli were presented, generalization is consistent irrespective of sampling explanation.

To assess the weight of evidence for any effects of sampling manipulation and presentation order we compared model fits for three generalized linear models. Across all models, counts of positive responses (indicating that the test item is an example of the target category) were modeled

[1]Note that there is still some tightening in the WEAK condition, suggesting that participants did not entirely abandon the apparent default assumption of strong sampling; however, the tightening was less than in the STRONG condition.

for each individual and stimulus using a beta-binomial distribution and a logistic link function.[2] The BASELINE model includes two predictors common to all three models. Firstly, it includes the individual's previous response to the test stimulus, which (per the Bayesian generalization model) is intended to model the dependence of an individual's posterior beliefs (measured after the second training phase) on their prior beliefs (measured after the first training phase). Secondly the BASELINE model includes an interaction with the test stimulus, allowing us to quantify the support for an effect that diminishes with increasing similarity between test and training stimuli. The +MANIPULATION model extends the BASELINE model with predictors for the sampling explanation and presentation sequence (and an interaction between the two). Lastly, the FULL model extends the +MANIPULATION model with all possible interaction terms. Completing the interaction terms in this way allows us to capture an effect of our experimental manipulation that is stimulus specific, in line with the predictions of the Bayesian generalization model (illustrated in Figure 1). Models were fit to response data from the second testing phase, for the 15 (of 19) stimulus values which lay outside the range spanned by the training stimuli.[3]

While alternative models involving different combinations of parameters may be equally plausible a priori, our goal is not to find the best fitting model for its own sake. Rather, because the central question of concern is whether our sampling manipulation was effective when it was made explicit before or after the training data was observed, the main comparison of interest is between the BASELINE model and the +MANIPULATION model. However, whereas the +MANIPULATION model represents something of a minimal model (assuming our experimental manipulation has some effect), the FULL model is able to more closely capture the predictions of the Bayesian generalization model. The aim in including both the FULL model and the +MANIPULATION model is to allow us to use the better fitting alternative in order to estimate the magnitude of the effect of interest.

Table 1 shows leave-one-out cross-validation information criteria (LOOIC) for each of the models considered. LOOIC has several advantages over simpler information criteria such as AIC and DIC, and is computed using posterior samples of the model parameters and gives a pointwise measure of the models' ability to predict out-of-sample observations (Vehtari et al., 2017). In keeping with AIC and DIC, LOOIC incorporates a penalty for model complexity in the form of an estimate of the effective number of model parameters. After incorporating appropriate penalties for the complexity of the +MANIPULATION model, in practical terms the model has equivalent predictive accuracy to that of the BASELINE model ($LOOIC_{\text{diff}} = 3.5$, $SE = 2.4$). Further, we find that the complexity involved in the additional interaction terms of the FULL model is not warranted in terms of the gain in predictive accuracy over the +MANIPULATION model ($LOOIC_{\text{diff}} = 128.3$, $SE = 16.9$). The lack of support for the FULL model is not wholly in keeping with the qualitative

[2] Our response count data is overdispersed – participants made four separate binary responses for each test stimulus and together $0/4$ and $4/4$ account for 92% of the counts observed. In this case the beta-binomial model, which allows the binomial rate to vary across observations is a robust alternative to the (fixed rate) binomial model (Gelman et al., 2014).

[3] Models were fit using the brms package version 2.13.0 (Bürkner, 2018) in R (version 3.6.3). All models contained a random intercept term for each participant. A weakly informative Gaussian prior ($\mu = 0, \sigma = 5$) was placed on all (logit scale) coefficients. Where it was included, the predictor for baseline response count was scaled onto the range $[-.5, +.5]$. All other predictors were coded as categorical variables.

predictions of the Bayesian generalization model (as illustrated in Figure 1). However, given that our experimental manipulation was able to elicit only a modest overall change in generalization between test phases, the scope for exemplar-specific interactions was necessarily limited.

Despite the lack of a clear distinction between the BASELINE and +MANIPULATION models, it remains plausible that there is a modest interaction in one cell of the 2×2 design manipulating sampling explanation and presentation sequence. We therefore used the +MANIPULATION model to test the hypothesis that people's responses differed according to the sampling explanation given, testing separately for the BEFORE and AFTER conditions. When people are given the sampling explanation before the second training phase, parameter estimates drawn from the posterior distribution suggest that they are 2.5 times more likely to endorse a stimulus as belonging to the target category if given the WEAK (rather than STRONG) sampling explanation (95%CI: $1.5\times$ to $4.4\times$ more likely, $BF_{10} = 20:1$).[4] In contrast, when people are given the sampling explanation only after the training items are presented there is strong evidence that the sampling explanation has no effect (95%CI: $1.6\times$ less likely to $2.1\times$ more likely, $BF_{01} = 20:1$).

Taking into account the modest effect of the sampling manipulation in the BEFORE condition, we interpret the pattern of results from our analyses as evidence in favor of an *encoding only* explanation. In light of this conclusion, the results for the AFTER condition are instructive about peoples' default sampling assumptions. Inspection of Figure 4 shows that, in the AFTER condition, generalization patterns in both sampling assumptions groups resembled the STRONG BEFORE condition. Consistent with previous work (Hayes, Navarro, et al., 2019; Ransom et al., 2016), this suggests that strong rather than weak sampling is a more likely default assumption about the data generation process in many experimental settings.

To test the generality of these findings we consider a different context with different sampling assumptions with a qualitatively different effect on generalization. For this we turn to our second experiment.

Experiment 2

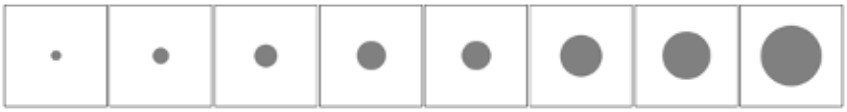
The aim of Experiment 2 was similar to that of the previous experiment – to test encoding and retrieval accounts of sampling assumption effects by varying the timing of these assumptions relative to sample data. In this case however, our focus compared different kinds of sampling frames (rather than strong and week sampling) and the task involved generalizing a novel property rather than judging category membership.

In this task, participants were presented with a training sample from a single category (small extra-terrestrial rocks) with a novel property (the valuable mineral "plaxium"). They were then asked to make inferences about the extension of this property: whether it was likely to be found in other rocks of various sizes. As in Experiment 1, generalization judgments were made first after only a small number of sample observations and again after a larger sample was observed. As in

[4] Here the Bayes' factor measures support for a non-zero change in the log-odds of a positive response, and was calculated using the Savage–Dickey density ratio method (Wagenmakers et al., 2010).

Figure 5

Stimuli used in Experiment 2



Note. Rocks of all sizes (R1-R8) were presented during the test phase. In all conditions, only R2 and R3 were presenting during training.

Hayes, Banner, Forrester, et al. (2019), we varied whether people were given cover stories implying PROPERTY sampling (under which participants saw any rocks containing plaxium and should lead to tighter generalization with increasing sample size) or CATEGORY sampling (under which participants saw only small rocks and should not lead to tighter generalization with increasing sample size). Like Experiment 1, we varied whether people were given the sampling cover story BEFORE or AFTER the second set of samples. According to the *encoding account*, tightening of generalization under property sampling should only be found in the BEFORE condition, while the *retrieval account* predicts that tightening should be evident regardless of the timing of the instructions.

Method

Participants

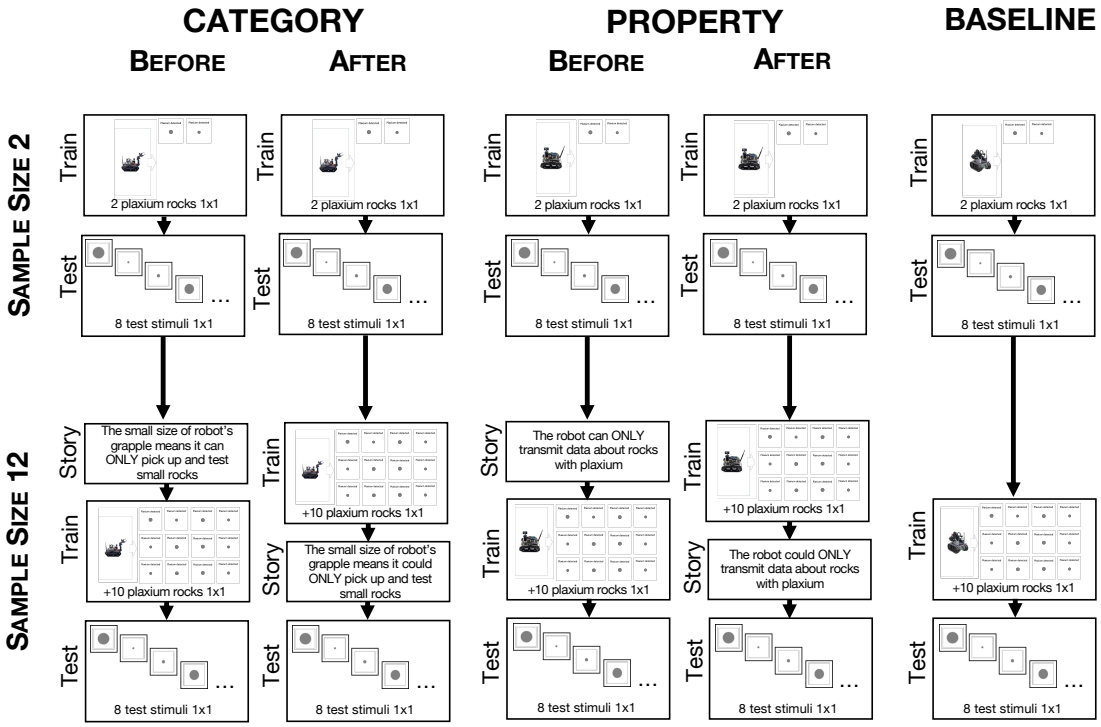
We recruited 634 people via Amazon Mechanical Turk who were each paid $1.20 for 10-15 minutes participation. 41% were female, with age varying between 19 and 87 (median: 34 years). Participants were U.S.-based, met a minimum approval rating of 95%, had completed at least 1000 HITs, and and had not previously completed an experiment involving the plaxium cover story. Participants indicated informed consent via an online consent form. The research was approved by the School of Psychology Human Research Ethics Subcommittee of the University of New South Wales (approval code: 3085). All procedures were performed in accordance with the institutional guidelines.

Stimuli

Participants in all conditions saw identical samples of evidence about rocks that had a novel property ("plaxium"). As Figure 5 shows, the rocks consisted of grey circles differing only in their diameter in pixels (R1 = 15, R2 = 24, R3 = 34, R4 = 43, R5 = 53, R6 = 62, R7 = 71, R8 = 90). During the test phase people saw all eight rock types, while in the training phase in all conditions people only saw two of the smaller ones (R2 and R3).

Figure 6

Design for Experiment 2



Note. Our 2×2×2 design varied Sample Size within-subject and Sampling Explanation and Presentation Sequence between-subjects. All participants began by seeing two individually-presented exemplars followed by a generalization task to novel stimuli. Those in the CATEGORY condition were then given a cover story in which the subsequent 10 items were chosen because only small rocks would fit into the robot's small collecting claw. Those in the PROPERTY condition were told that the items were selected because the robot had detected plaxium. In the BEFORE condition, the cover story was given before seeing the 10 new items; in the AFTER, it came after. In all conditions the experiment ended with a repeat of the generalization test.

Design and procedure

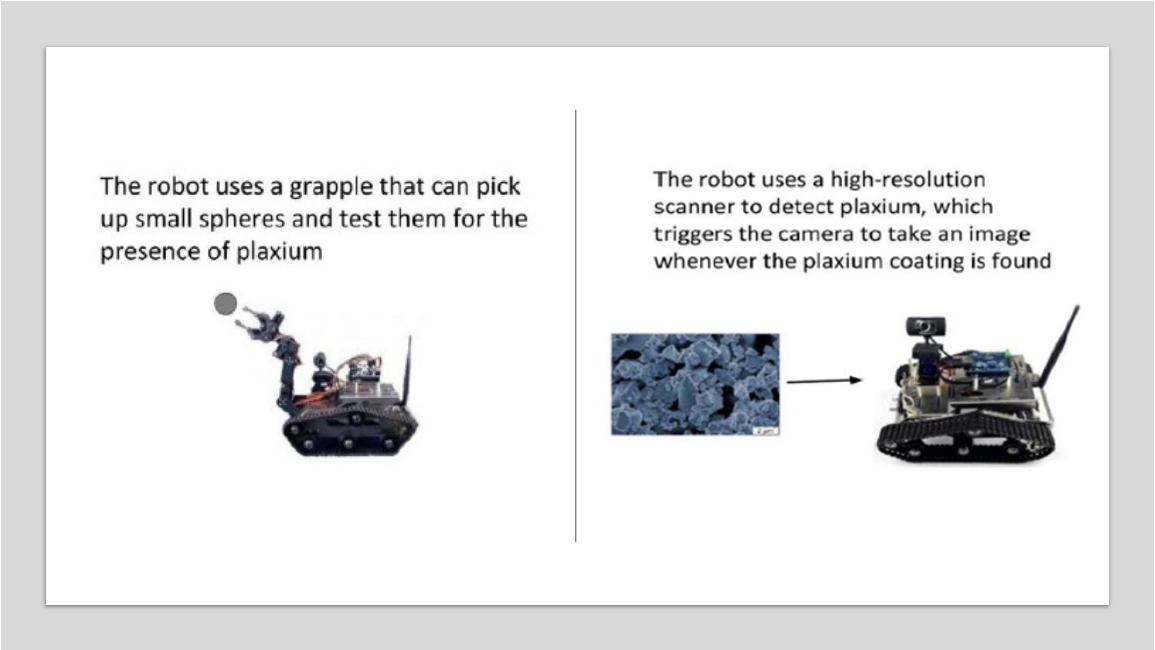
As in Experiment 1, this experiment employed a $2 \times 2 \times 2$ mixed factorial design which manipulated two factors (Sampling Explanation and Presentation Sequence) between-subjects while another (Sample Size) varied within-subject. This time, as a comparison, we also included a NO STORY condition in which participants never saw a cover story at all.[5] This resulted in five conditions which people were randomly allocated to: CATEGORY BEFORE, CATEGORY AFTER, PROPERTY BEFORE, PROPERTY AFTER, and NO STORY.

Sample Size The experiment involved two rounds of testing in all five conditions. The first

[5]We also manipulated between-subjects whether sample items accumulated on-screen during training. This manipulation had no impact on our main findings and for simplicity we collapse all of this data together.

Figure 7

Screenshots of Instructions used for Category and Property Sampling Conditions.



(Size 2) occurred after a training phase involving two training examples (one of R2 and one of R3), and the second (Size 12) after seeing ten more (all either R2 or R3). All items during training and test were presented in random order.

Presentation Sequence This between-subjects manipulation varied when the sampling cover story was presented in relation to the second training set. People in the BEFORE condition were told the cover story (CATEGORY or PROPERTY, described below) *before* viewing the second set of training items, while people in the AFTER condition were offered the identical explanation (with verb tenses adjusted) only after all training items had been presented. Presentation sequence was not manipulated for people in the NO STORY condition, who were not told a cover story at all.

Sampling Explanation The other between-subjects manipulation varied the details of the cover story explaining how the data in the second training phase were generated. The initial training phase, however, was identical for all participants. No explanation was given for how the exemplars were chosen. People were told only that they had been placed in charge of a robot probe sent to explore the alien planet of Sodor, which had spherical rocks that varied in size. Their task was to use the robot probe to establish which spheres had plaxium and which did not. After reading these instructions, participants completed a two-part comprehension check which they had to answer correctly to proceed to the main task. If they answered either of the questions incorrectly, they were taken back to the start of the instructions. After correctly answering the comprehension questions, they were told that the probe would start transmitting data shortly and there would be a pause every few transmissions to ask for their guesses about which spheres had plaxium coating.

In each training phase trial participants clicked a button to reveal a rock sample, the sample item appeared on the screen, and then after 3 seconds a button appeared and conveyed the message "The probe has detected a sphere: Click here to view" after which another sample item appeared. In the second training phase people were given one of two different cover stories explaining how the items were selected. As Figure 7 shows, people in the CATEGORY condition were told that only small rocks could be sampled because only these would fit into the robot's small collecting claw. Those in the PROPERTY condition were told that the rocks they were shown were the first that had shown a positive result when photographed using a plaxium-sensitive camera. Participants in both conditions had to correctly complete a second series of comprehension check questions about these instructions before they could continue. Regardless of condition, by the end of the second training phase, all participants had observed the same sample: 12 small rocks (R2 or R3), all with plaxium.

Generalization test Immediately after both the first and second training phase, participants in all conditions performed the same generalization test. They were shown the eight rocks (R1 through R8) in random order and were asked asked to rate the likelihood that each test item had plaxium (1 = Very Unlikely, 10 = Very Likely).

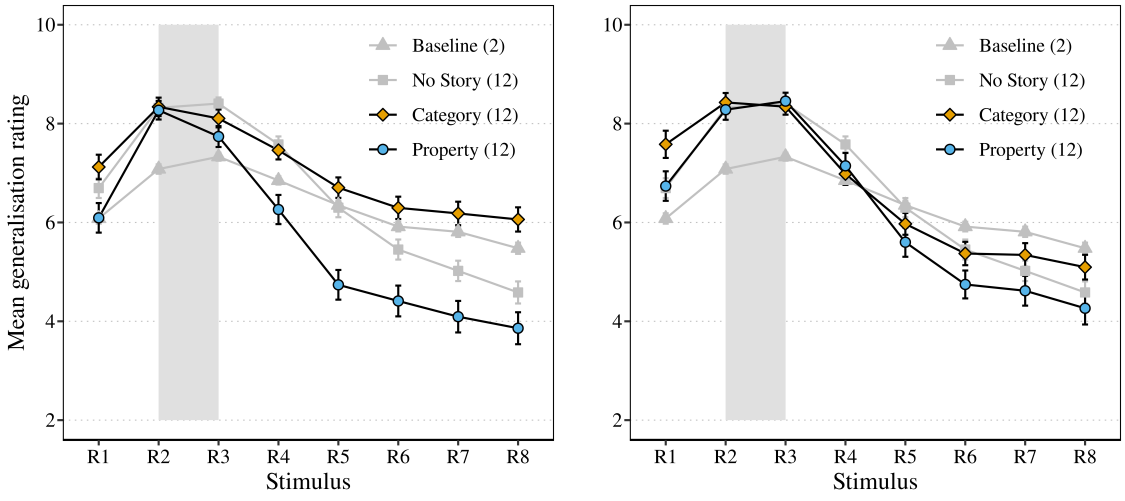
Results

As for the previous experiment, we began our analyses by constructing aggregate generalization curves, shown in Figure 8. The baseline curve (grey triangles) represents mean generalization ratings (collapsed across participants in all conditions) made when people had only seen two training items (R2 and R3) for which no sampling explanation was given. The remaining curves represent people's generalization ratings after seeing the additional ten examples. In the BEFORE condition, a comparison of the different gradients suggests a marked and qualitatively different effect of the additional training items on the pattern of generalization, depending on the sampling explanation offered. In the CATEGORY condition, where people were told that the restricted range of observations was an artefact of the sampling process (i.e. the probe could only gather small rocks) people increased their willingness to generalize the property in question across the full range of test items. In the PROPERTY condition in contrast, where people were told they would only see positive examples of the property in question, they appeared significantly less willing to generalize the property outside the narrow range of training items. When comparing generalization gradients in the AFTER condition however, the effect of sampling explanation is less evident. When the method of sampling was explained only after seeing the additional training items, people appeared to have tightened their generalization regardless of which explanation they were given.

Taken at face value, the generalization gradients shown in Figure 8 appear to rule out a *retrieval only* explanation. This suggests that the learner's beliefs about the generative process behind the data at the time it is first observed affect the way that the data is encoded, which in turn influences subsequent generalization – even if the learner comes to revise these beliefs by the time that the data are recalled. To quantify the evidence in favor of this finding, we first compared how well each of four different regression models fit the generalization ratings given during the

Figure 8

Property Generalization as a Function of Presentation Sequence, Sampling Explanation and Sample Size



(a) BEFORE. (b) AFTER.

Note. The figure illustrates people's performance on the property induction task featured in Experiment 2. Participants rated the likelihood that each test item had the property "plaxium". The graphs show mean ratings for each of the test stimuli, and error bars indicate standard error of the mean. People's performance after seeing two training items with no sampling explanation given (grey triangles) is contrasted with their performance after seeing an additional 10 examples for which a sampling explanation was provided (black lines). As an additional control, one group of participants saw the additional 10 items without any explanation as to how the items were sampled (grey squares). (a) When the sampling frames explanation was given prior to the presentation of the final 10 examples (BEFORE condition), people who were told that training items were selected because they had the property in question tightened their generalizations, while people who were told that the samples reflected the category from which they were sampled (small rocks), were more willing to generalize. (b) In contrast, when the sampling frames explanation was given only after all training stimuli were presented (AFTER condition), the sampling manipulation appeared to have a reduced effect, with people tightening their generalization to a similar degree in category and property frames conditions.

second test phase, for all novel stimuli (i.e. excluding R2 and R3 which were seen in the training phase). To avoid the pitfalls of analysing ordinal data using metric models (see Liddell & Kruschke, 2018), responses (which were given on a scale of 1 to 10) were modeled using a cumulative-logit response function (Bürkner & Vuorre, 2019). Differences in the response function notwithstanding, the four models have the same structure and justification (and were fit with the same software, prior specification, and approach to predictor scaling) as those used in our analysis of Experiment 1.

Table 2 shows leave-one-out cross-validation information criteria (LOOIC) for each of the models considered. Taking both model fit and complexity into account, the analysis reveals that the FULL model gives the best account of the data from Experiment 2. Fits of the FULL model to people's responses are illustrated in Figure 9 on an item by item basis (for all novel test items). Despite

Table 2

Comparison of how well three different cumulative-logit regression models capture the responses from the second test phase of Experiment 2 (lower LOOIC indicates better fit).

Model	Model performance				
	LOOIC	SE	Contrast	$\text{LOOIC}_{\text{diff}}$	SE_{diff}
1. BASELINE	14057	117	–	–	–
2. +MANIPULATION	14060	117	LOOIC_{2-1}	3.4	6.1
3. FULL	13985	121	LOOIC_{3-2}	-74.8	32.0

Note. The FULL model is preferred suggesting both that the experimental manipulation was successful, and that the way that people encode observations may be impacted by the sampling assumptions that they hold at the time. See Table 1 for details of the predictors involved in each model.

some limitations of the cumulative-logit response model in capturing responses in the middle of the rating scale (evident in the top row for items R7 and R8), our model analyses lends overall support to our previous finding that the effect of sampling explanation differed depending on when it was given, and casts further doubt on a *retrieval only* account.

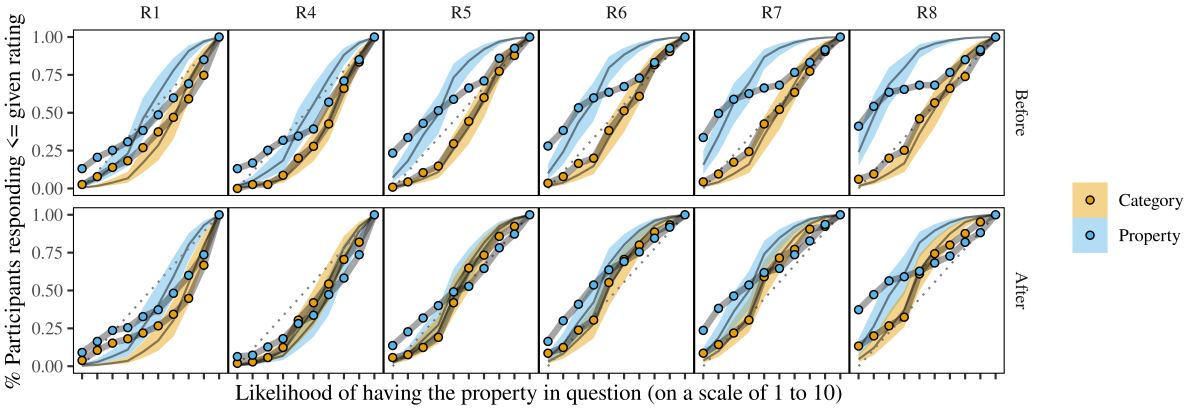
While this evidence in favor of the FULL model represents a contrast to the findings from Experiment 1 (where the +MANIPULATION and BASELINE models were favored), there are differences in the structure of the stimuli and the sampling assumptions involved that may help to explain the contrast. In Experiment 1, the Bayesian generalization model predicts a reduced (or equal) willingness to generalize to novel exemplars between the first and second test phases, regardless of the test stimuli involved. In practice, given that a modest overall reduction in generalization was observed in both the STRONG and WEAK sampling conditions, it is not surprising that the increased fit of the FULL model failed to compensate for the additional model complexity.

In Experiment 2 however, the Bayesian model predicts that category sampling has a qualitatively different influence on generalization for stimuli R1 (a "small rock" that is admissible under the category sampling frame yet is never observed in the learning phase) compared to stimuli R4–R8 ("large rocks" which cannot be seen under the category sampling frame). This qualitative difference due to the structure of the stimuli helps to explain why the additional complexity of the FULL model in terms of exemplar-specific interactions was justified in Experiment 2 (see Appendix A for further details regarding the Bayesian model predictions across these two types of stimuli). However, while these nuances are relevant to the task of selecting the appropriate model to use in order to quantify support for the hypothesis of interest – whether our sampling manipulations have an effect when presented before or after the training stimuli – the specific details of the interactions involved are not our focus here.

To further quantify the evidence for our findings and to attempt to distinguish *encoding only* and *encoding and retrieval* accounts, we used the posterior distribution for the FULL model obtained during the model fitting exercise to examine parameter estimates and test hypotheses using

Figure 9

Fits of the Best Fitting Model to Generalization Ratings by Stimulus and Sampling Explanation



Note. Each plotted curve represents the percentage of people responding with a rating less than or equal to a given rating. Shaded curves represent Bayesian fits drawn from posterior distribution of the (FULL) fitted model, while plotted points represent the empirical data. When people were given the sampling explanation before the second phase of training there was a tendency to avoid the middle of the rating scale. This can be seen for those stimuli most dissimilar to the training items (R6 to R8 in the top row). The cumulative response model struggles to fit these bi-modal distributions. Overall, the plots reveal greater separation between the CATEGORY (yellow) and PROPERTY (blue) conditions in the BEFORE condition (top row) than the AFTER condition (bottom row). This suggests that the sampling assumptions people make when observations are first encountered may impact later generalization by changing how observations are encoded, and that while revised assumptions at the point of retrieval may too play a role, the impact is greatly reduced.

the Savage–Dickey method (Wagenmakers et al., 2010). In the BEFORE condition we found decisive evidence in favor of an effect of sampling explanation on generalization ($BF_{10} > 10^3 : 1$). In contrast, the same test with respect to the AFTER condition yielded substantial evidence in favor of the null ($BF_{01} = 3.23 : 1$). Directly comparing the magnitude of the two effects revealed strong evidence in favor of a difference ($BF_{10} = 29 : 1$). Next, we separately compared the responses of people in the two AFTER conditions to those from the NO STORY condition (Figure 8, grey squares), where no sampling explanation was given. The comparison revealed substantial ($BF_{01} = 7.6 : 1$) to strong ($BF_{01} = 29 : 1$) evidence in favor of the null (i.e. no difference, consistent with an *encoding only* account), for property sampling and category sampling respectively. Taken together these analyses suggest the same qualitative result as in Experiment 1: people's ability to incorporate new sampling assumptions into their reasoning once the data is encoded may be limited at best. We turn now to discuss some general implications of our findings.

General Discussion

A view widely expressed in models of property induction (e.g., Heit, 1998; Osherson et al., 1990; Sloman, 1993), stimulus generalization (e.g., Estes, 1950; Shepard, 1987) and category learn-

ing (e.g., Medin & Schaffer, 1978; Nosofsky, 1986) is that generalization from a sample of items to a novel item is governed by the comparison amongst the items without explicit regard for how the sample was constructed. However, our results add to the mounting body of evidence suggesting that sampling assumptions mediate generalization behavior (Hayes, Banner, Forrester, et al., 2019; Hayes, Navarro, et al., 2019; Hendrickson et al., 2019; Navarro et al., 2012; Ransom et al., 2016; Xu & Tenenbaum, 2007). Critically, our work extends previous research by showing that the sampling manipulation had a substantial effect only when it was presented *before* people viewed the training sample. This finding supports an *encoding only* account: generalization behavior appeared to be governed by the sampling assumption which held when the data were first encoded, and did not change even if the explanation given after encoding contradicted the initial (default) assumption (which must have been the case for at least one of the two sampling explanations in each experiment). To our knowledge, our work is the first to demonstrate that generalization is dependent not only on sampling assumptions but also on *when* those assumptions are first brought to bear.

Given that we find little to no effect of the sampling manipulation in the AFTER conditions, we can reasonably infer that whichever sampling mechanism was implicit during the training phase was the assumption that guided generalization. On this basis, our results are also informative about what assumption people adopt when there is no explicit guidance regarding the sampling mechanism, as is the case for most previous studies of inductive inference. To the extent that a pragmatic or pedagogical interpretation seems reasonable when dealing with socially generated data like category labels, the strong sampling bias exhibited in the context of Experiment 1 is unsurprising. A similar bias is also evident in the NO STORY condition in Experiment 2, where no sampling explanation was provided. Despite the lack of an explanation, people still tightened their generalizations with increasing sample size, as occurs under strong sampling and property sampling. Across both experiments it appears that the absence of sample items outside the range observed is taken as evidence that the concept in question does not extend beyond it, *unless* a reason to discount this "absence of evidence as evidence of absence" is made explicit before the sample items are encountered. This is consistent with the findings of Ransom et al. (2016), who found that strong sampling was the default, and people were only willing to go against it when given an explicitly random cover story.

How are sampling assumptions represented and what do they represent?

From a Bayesian perspective, our results suggest that the likelihood function – the key component in assessing the evidentiary weight of data – is evaluated at the time of encoding, and not re-evaluated when a generalization decision must be made. This distinction is important as it has implications for the role that representation plays in generalization. Under a *retrieval only* account, evidence in favor of competing hypotheses need only be represented when a target generalization is being assessed, and only for the duration of the decision process. By potentially decoupling the long-term representations of data from the representations required for in-the-moment generalization, such a scheme would admit considerable flexibility allowing the two concerns to be treated independently. However, since our finding supports an *encoding* account instead, this suggests a

greater dependency between the nature of generalization and the initial encoding format.

How might such a dependency be realized? One possibility is that both factors (the initial data and the sampling assumption) are encoded together, such that generalization requires accessing both. Another possibility is that the sampling assumption affects how the data are encoded, so that the representation of the observed data is shaped by that assumption but the assumption itself is not encoded. Future investigations might test this conjecture by attempting to dissociate sampling assumptions from encoding. For instance, one might explore whether or not the effects of a sampling manipulation remain when participants are unable to reliably recall the sampling manipulation used. Such a test might help to resolve an apparent contradiction between our finding and Fiedler's (2012) view that people draw misleading inferences because they are largely blind to matters of sample construction. If people fail to retain an explicit memory for how information was sampled, but nonetheless retain an encoding of the sample shaped by the original assumption, then there need be no contradiction. Certainly, the lack of a sampling assumptions effect in the AFTER conditions across our two studies is consistent with the general finding that people find it difficult to retrospectively apply information about data generation processes when making judgments and inferences (Fiedler, 2012).

Our findings suggest challenges for current models of inductive inference. How do we identify which elements are represented at encoding and which at retrieval only? How do we identify how different sampling assumptions lead to different representations? For Bayesian models, one possibility is to use causal Bayes nets to capture sampling assumptions as well as the causal structure and mental models people use in reasoning tasks such as these (Bareinboim et al., 2014; Kemp et al., under review; Sloman & Lagnado, 2015). The view of sampling assumptions as something akin to causal explanations is consistent with Murphy and Medin's (1985) notion that the causal mechanisms contained in people's theories are the very means by which feature correlations are represented.

In this paper we have used Bayesian cognitive models because they usefully highlight the general computational structure of the inference problem as well as the principles by which different kinds of sampling license different inferences. However, both the problem and the insight are more general. For instance, connectionist models might be able to implement different kinds of sampling assumptions as well. There already exist feature-based connectionist models that can account for a wide range of phenomena in the property induction literature (Rogers & McClelland, 2004; Sloman, 1993). By controlling how individual features in the data layer are connected to inner layers (in line with causal theories), such models might be extended to allow sampling assumptions to be captured at the feature level. In certain contexts, such as Experiment 2 for example, it is not unreasonable to consider that while one feature of the data (e.g. the size of Sodor rocks) is effectively censored by the sampling mechanism, another feature (the property base rate) is uncensored and may be representative of the population of items more broadly. A connectionist network formed in this way has the potential to learn a representation of the concept in question where the likelihood of data is bound up in the network weights.

Although similarity-based models do not directly account for the sampling process, attempts to accommodate an encoding account of sampling assumptions by constraining the interpretation of free parameters may yield useful insight. For example, the similarity-coverage model of property induction (Osherson et al., 1990) uses a parameter α to weight the contribution of evidence directly from items in the sample (*similarity*) and indirectly via an encompassing category (*coverage*). The tightening of generalization with increased sample size that is characteristic of a strong sampling assumption can be captured if the weighting shifts in favor of similarity as sample size is increased.[6] Along similar lines, the Generalized Context Model, a prominent similarity-based model of categorization (Nosofsky, 1986), might also accommodate strong sampling by increasing the scale parameter c as sample size increases (for details see Hendrickson et al., 2019).[7] In both models, evidence from the sampling process can be incorporated by recasting key generalization parameters as encoding parameters tied to the representativeness of samples. Formalising such relationships requires identifying a basis for encoding the representativeness of a sample in a way that is distributed across the items in the sample and the concept being learned.

Broader implications

Our work connects with a question of long-standing interest about what exactly generalization *is*. For instance, Razran (1949, p. 362) held the view that "all effects of generalization are generated during tests of generalization." In an experiment using unfamiliar words, the generalizations of people given conflicting meanings before and after training (conditioning) were compared with those of people given meaning only after training. The results revealed that meanings given prior to training had no effect on generalization. Early models of stimulus generalization (e.g., Bush & Mosteller, 1951; Estes, 1950; Hull, 1943) contrasted with this "retrieval only" account by suggesting that generalization gradients are determined by the associations formed when stimuli are first represented (see Lovibond et al., 2020, for a recent review). However, because the early representations were assumed to be based on "stimulus elements" or low level features, these models established the concept of "theoryless" learning where data is stored in a veridical fashion, unbiased by the learner's knowledge and assumptions.

Our finding contrasts with both of these ideas, suggesting that generalization is shaped from the moment of encoding because from the outset learners actively interpret data and form representations of concepts that go beyond raw data. As a computational strategy this form of learning has its advantages: pre-computing the evidentiary weight of data on the basis of the learner's theories and assumptions has the potential to reduce memory load and support speeded generalization at decision time. However, these benefits must be traded off against the risks of reasoning on the basis

[6] This has some intuitive appeal. In statistical terms, the coverage component can be viewed as a regularization term that reduces overfitting by recruiting background knowledge when the available sample is limited. As the sample becomes more representative of the concept of interest (as sample size grows), the risk of overfitting diminishes and the contribution of the regularization term can reasonably be reduced (towards zero in the limit).

[7] Alternatively, interpreting the GCM as kernel-based probability density estimation (Ashby & Alfonso-Reese, 1995) suggests that strong sampling can be captured in terms of decreasing kernel width with increased sample size.

of stale or incorrect theories. Intuitively, the balance of these risks might shift across the developmental trajectory, although much research remains to be done to determine whether this is actually the case. To what extent the trade-off between "theoryless" and "theory-influenced" representations changes across the lifespan or might be under learner's control is thus an open question.

Broadly construed, our findings have implications for effective pedagogy as well. It is known that learners benefit by assuming that their teacher is selecting the most informative examples possible given the learner's current beliefs. Such reciprocal assumptions can lead to a highly leveraged form of generalization in which concepts can quickly be acquired from minimal input (Shafto et al., 2014). Under the idealised account of pedagogical learning, people's inferences should not depend on when the sampling process becomes apparent. However, our results suggest both that it is important for the teacher to explain the sampling process as early as possible in order to avoid distortions of data encoding that cannot be easily revised.

In a similar way our finding has implications for how people process misinformation and corrections to misinformation. Ransom et al. (2017) found, for example, that people can use truthful but limited data in their efforts to mislead others by attempting to manipulate their counterpart's sampling assumption. An encoding account of sampling assumptions suggests that subsequently learning that an information source was biased may not be sufficient to correct the bias. Our work thus connects with the literature on the continued influence effect, a well-established finding that retracting misinformation does not eliminate its influence (Connor Desai et al., 2020; Ecker, Lewandowsky, Swire, et al., 2011; Johnson & Seifert, 1994) The selective retrieval account of continued influence holds that old and new information compete at the point of retrieval and that previously encoded misinformation may be inadequately suppressed (Ecker, Lewandowsky, Swire, et al., 2011). In contrast, the model updating view holds that revised information is poorly encoded when it is less coherent with people's mental model of how the original data/event came about (Ecker, Lewandowsky, & Apai, 2011; Johnson & Seifert, 1994).

Our findings argue against the selective retrieval account of continued influence effects. Although they are in some ways consistent with the model updating account they also offer an alternative explanation: if people are encoding data in such a way that it cannot be disentangled from their theory at the time, interpreting that data under a new theory may be extremely difficult. An interesting test to distinguish these explanations would be to employ our experimental design in a context where the sampling assumption provided after seeing the data is a *better* explanation than the default assumed beforehand. A model updating account would predict that the explanation would "correct" the assumption held at the time of encoding because it is more coherent, whereas our explanation would predict that it should not.

Conclusion

Overall, our research advances our understanding of the impacts of sampling assumptions on categorization and inference - showing that such assumptions primarily affect the way we encode new evidence rather than the way we retrieve previously learned evidence. This work shows that

memory, sampling processes, and generalization are intertwined in ways that are still not fully understood, and that generalization is determined in part by how information was encoded in the first place. By manipulating when different information is available and under what circumstances, we have begun to illuminate this complex relationship.

Appendix A: Mathematics of sampling assumptions

Computational models of the role of sampling assumptions in generalization have generally adopted a Bayesian approach. Tenenbaum and Griffiths (2001), for example, proposed a model of inductive reasoning where the learner's task is to infer the probability that a novel item y is an instance of some concept C.[8] This probabilistic decision is assumed to be guided by both the data observed d, and an assumption about how this evidence is generated, s. Formally this can be represented as:

$$P(y \in c|\mathbf{d},s) = \sum_{h \in \mathcal{H}_c : y \in h} P(h\,|\,\mathbf{d},s) \tag{1}$$

where h is a specific hypothesis about the true extension of C drawn from a set of alternative hypotheses $\mathcal{H}_C$ considered by the learner.[9] Equation 1 holds that the evidence in favor of the item y being an instance of the concept C is effectively averaged over all those hypotheses entertained by the learner under which the relationship holds. In this way, the equation captures an assumption that underpins the Bayesian view of generalization – namely, that generalization reflects the learner's uncertainty about the extent of a given concept (Shepard, 1987). A straightforward application of Bayes' rule captures how the learner weighs the evidence for each hypothesis considered:

$$P(h\,|\,\mathbf{d},s) \propto P(\mathbf{d}\,|\,h,s)P(h). \tag{2}$$

That is, having made an assumption s about how the observed data $\mathbf{d}$ are sampled, the probability that the learner assigns to a candidate hypothesis h is a joint function of the likelihood of observing the sample data given the hypothesis and the sampling assumption, $P(\mathbf{d}\,|\,h,s)$, and the prior probability of the hypothesis, P(h).[10]

Within this framework, it is the likelihood function that is crucial for capturing the effects of different sampling assumptions. The likelihood functions corresponding to the four sampling assumptions featured in the main text share a common form, each placing different restrictions on the data that may arise. This relationship is captured in Figure 10, which illustrates the sampling

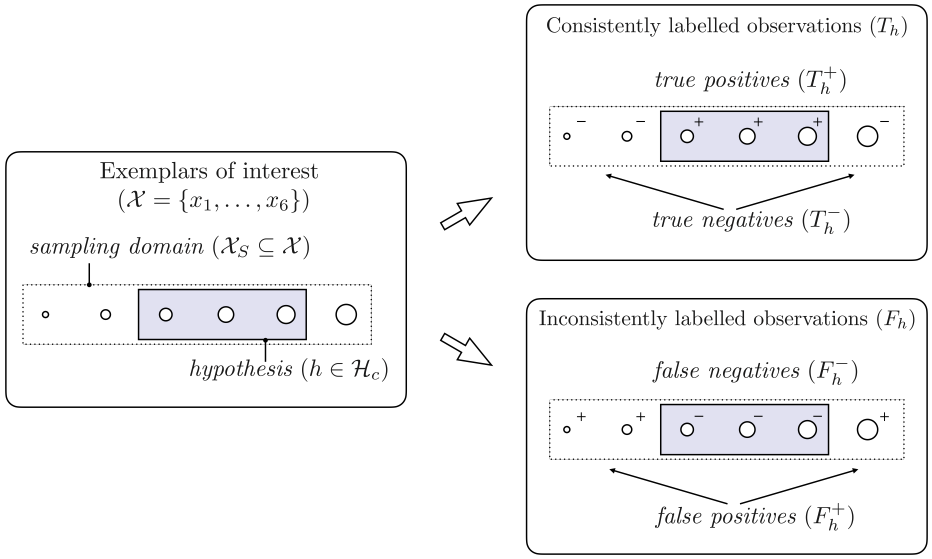
[8] Importantly for our present purposes, the notion of *concept learning* modeled by Tenenbaum and Griffiths (2001) is flexible enough to accommodate both the single-category learning task used in Experiment 1 (where C represents the category to be learned), and the property induction task used in Experiment 2 (where C represents those exemplars sharing the target property).

[9] Throughout this appendix we assume that hypotheses and exemplars are each drawn from a discrete space. In general, the appropriate equations for the continuous case can be obtained by substituting integration for summation.

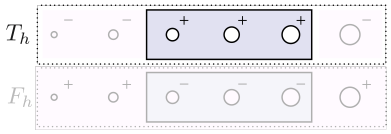
[10] For current purposes, this formulation assumes that the learner entertains a single sampling assumption independent of the concept in question (i.e. $P(h,s) = P(h)P(s) = P(h)$), which we presume was given to them by a cover story describing the generative process.

Figure 10

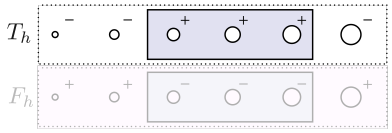
An Example of Reasoning About Observations on the Basis of Sampling Assumptions



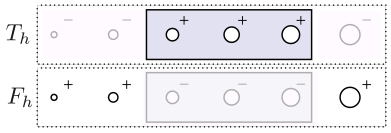
(a) Range of possible observations.



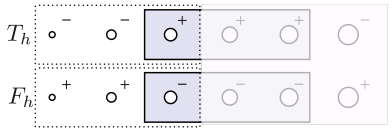
(b) Strong sampling.



(c) Weak sampling.



(d) Property sampling.

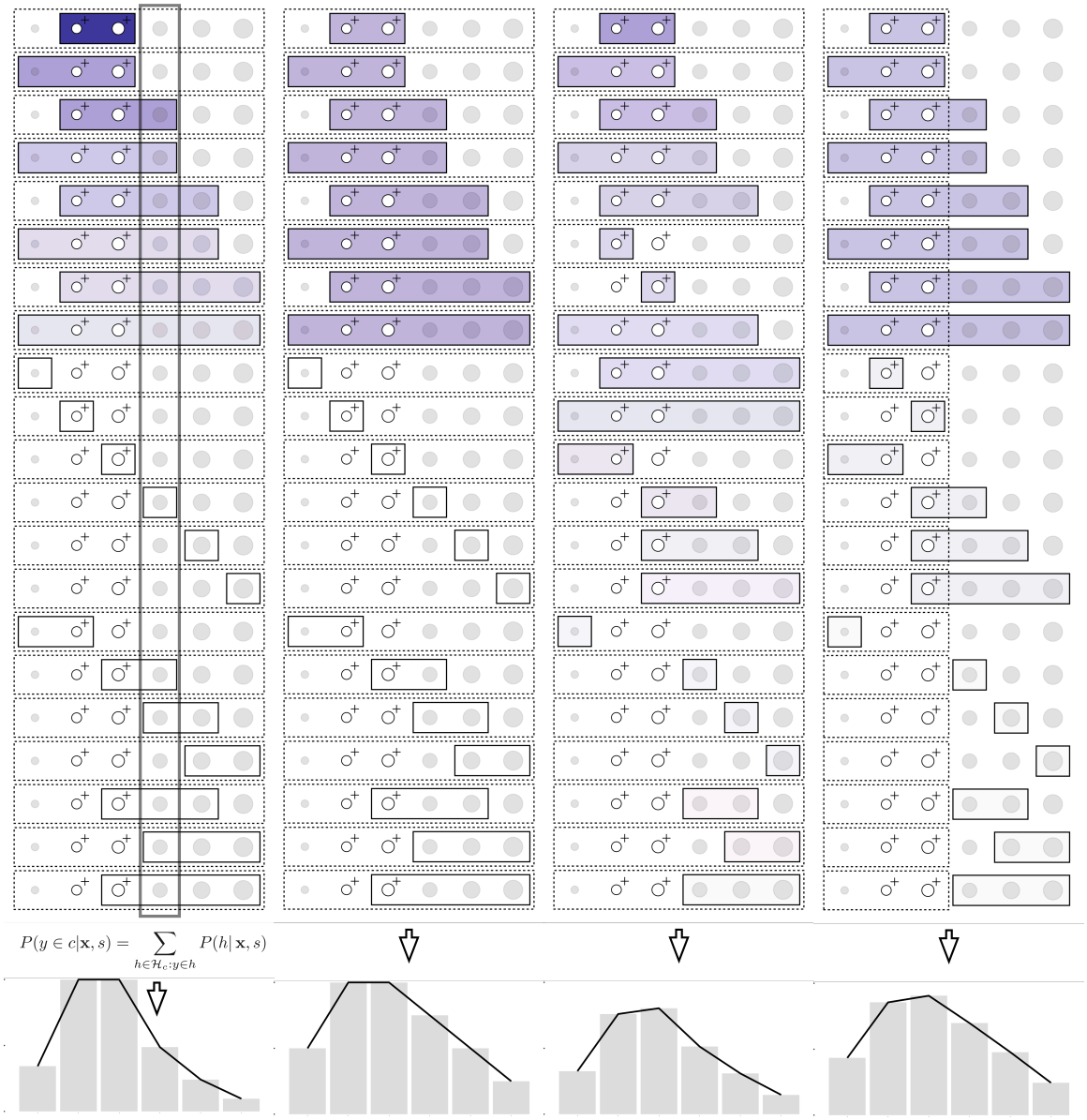


(e) Category sampling.

Note. The learner's task is to decide which of the exemplars of interest (six circles varying only in size) are an instance of some concept c. Observations are assumed to be drawn from the sampling domain and labelled according to whether or not the sampled item is an instance of the concept in question. (a) The learner entertains a particular hypothesis that three of the six items are instances. Under this hypothesis, and in the absence of any more specific assumptions, the learner can expect to see any of the observations shown in the two right hand panels. (b) If instead, the learner makes a *strong sampling* assumption, then only the three *true positives* shown are to be expected. Any other positive observation should invalidate this hypothesis. (c) Under *weak sampling*, only the six consistently labelled observations are to be expected given h; any inconsistently labelled observation should invalidate it. (d) Only positive examples of the concept are expected under *property sampling*; but inconsistently labelled examples will not invalidate the hypothesis outright. (e) *Category sampling* allows any of the four classes of observation, but only exemplars that fall within X_S are expected.

Figure 11

An Example of how Sampling Assumptions Impact Generalization Gradients



(a) Strong sampling. (b) Weak sampling. (c) Property sampling. (d) Category sampling.

Note. The panels depict four scenarios extending the example shown in Figure 10. Stacked boxes represent the learner's hypothesis space. From an initially uniform prior, the learner revises their beliefs (via Equation 2) on the basis of the two positive examples shown; darker shades indicate stronger belief, white indicates disconfirmed hypotheses. Generalization gradients (bottom panels) are computed via Equation 1; thus, bar height represents the sum of the hypothesis weights in the corresponding column. (a) Under strong sampling, the size principle dictates that smaller hypotheses consistent with the data are favored, hence the gradient is tighter around the exemplars observed. (b) In contrast, under weak sampling, data is used for disconfirmation only, so no tightening is evident. (c) To the extent that exemplars may be labelled incorrectly ($\theta = 0.8$ in this example) tightening around the data observed is less pronounced under property sampling than under strong sampling. (d) Under category sampling, observations are limited in range (dashed rectangles) and negative examples are possible; but because labelling is noisy ($\theta = 0.8$) belief revision depends on how the data and hypotheses overlap (note the three distinct shades corresponding to three levels of plausibility).

scenario in general (Figure 10a) and highlights the restrictions specific to each assumption (Figure 10b–e). In the general case, an exemplar x is assumed to be sampled from the sampling domain $\mathcal{X}_S$ containing some or all of the set of exemplars $\mathcal{X}$ in which the learner is interested. An observation $d \equiv \langle x, l \rangle$ corresponds to a sampled exemplar x and label $l \in \{+, -\}$ indicating whether the exemplar is an instance of the concept in question (belongs to a given category, or has a given property, for example). Allowing for the possibility that an exemplar may be labelled incorrectly, four different sets of observation become possible given an hypothesis h: *true positives* $T_h^+ \equiv \{x^+ \,|\, x \in \mathcal{X}_S,\ x \in h\}$, *true negatives* $T_h^- \equiv \{x^- \,|\, x \in \mathcal{X}_S,\ x \notin h\}$, *false positives* $F_h^+ \equiv \{x^+ \,|\, x \in \mathcal{X}_S,\ x \notin h\}$, and *false negatives* $F_h^- \equiv \{x^- \,|\, x \in \mathcal{X}_S,\ x \in h\}$. Taken together, true positives and true negatives ($T_h = T_h^+ \cup T_h^-$) represent observations that are logically consistent with the given hypotheses, while false positives and false negatives ($F_h = F_h^+ \cup F_h^-$) are logically inconsistent. The following likelihood function captures the general case:[11]

$$P(x^\pm \,|\, h, s) = \begin{cases} \dfrac{\theta P(x)}{\theta P(T_h^+) + \theta P(T_h^-) + (1-\theta)P(F_h^+) + (1-\theta)P(F_h^-)} & \text{if } x^\pm \in T_h \\ \dfrac{(1-\theta)P(x)}{\theta P(T_h^+) + \theta P(T_h^-) + (1-\theta)P(F_h^+) + (1-\theta)P(F_h^-)} & \text{otherwise} \end{cases} \tag{3}$$

where θ represents the probability that an observation is correctly labelled, and $P(x)$ represents the base rate of the observed exemplar adjusted relative to the sampling domain $\mathcal{X}_S$. The $P(.)$ terms in the denominator represent the combined base rate for any of the true positives, true negatives and so on – for example, $P(T_h^+) = \sum_{x^+ \in T_h^+} P(x)$.

Implementing the restrictions placed on observations shown in Figure 10 amounts to defining the sampling domain $\mathcal{X}_S$ and the noise parameter θ, and deleting the appropriate terms from the denominator of Equation 3. Substituting different likelihood functions into Bayes' rule (Equation 2) then yields different predictions about how the learner should generalize from given data. For example, under weak sampling all exemplars of interest are available for sampling ($\mathcal{X}_S = \mathcal{X}$) and are assumed to be correctly labelled ($\theta = 1$). Thus, because the denominator sums to 1 when only correctly labelled observations remain, the likelihood function simplifies to:

$$P(x^\pm \,|\, h, s_{\text{WEAK}}) = \begin{cases} P(x) & \text{if } x^\pm \in T_h \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

This reflects why it is a *weak* assumption: provided the observation is compatible with the hypothesis, then nothing further is assumed and the likelihood is simply the base rate for sampling the exemplar $P(x)$, which is independent of the hypothesis itself; otherwise, when the observation is incompatible with the hypothesis, the likelihood is zero. Taken together, Equations 2 and 4 reveal a consequence of this assumption, namely that data which is compatible with all hypotheses currently

[11] Here and elsewhere the notation $x^\pm$ implies that either a positively labelled exemplar x^+ or a negatively labelled exemplar x^- may be substituted consistently throughout the equation in which it appears.

entertained by the learner is rendered uninformative (as illustrated in Figure 11b).

In contrast, under strong sampling the same observation *is* informative, because the learner assumes a dependency between the data that was observed and the hypotheses under consideration. That is, because only correctly labelled positive examples are possible, the denominator in Equation 3 simplifies to $P(T_h^+)$. If we further assume that there is a uniform probability of sampling any exemplar, then the likelihood of seeing an exemplar x consistent with some hypothesis is inversely proportional to the size of the given hypothesis (denoted $|h|$).[12] We express this formally as:

$$P(x^+ \mid h, s_{\text{STRONG}}) = \begin{cases} \dfrac{1}{|h|} & \text{if } x^+ \in T_h^+ \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

By extension, if the data $\mathbf{x}^+$ represents a sequence of n independently sampled positive examples then this size is raised to the nth power $|h|^n$.

An important implication of Equation 5 is the "size principle"; under strong sampling, more specific hypotheses will generally be favored (i.e. will receive a higher posterior probability) over broader hypotheses, for a given sample of data (as illustrated in Figure 11a, for example).[13] Moreover, as more instances consistent with the hypothesis are observed, we should see "tightening" of generalization such that more specific hypotheses consistent with the data become even more likely.

Although the specifics differ, a broadly similar approach has been suggested for modeling generalization based on property and category sampling in property inference (Hayes et al., 2017; Hayes, Banner, Forrester, et al., 2019).[14] Equations 1 and 2 can be extended to the sampling frames scenario. In this case $P(h \mid \mathbf{d}, s)$ corresponds to the posterior probability of a hypothesis about the extension of the target property, and the sampling assumption s represents a "censoring function" for each sampling frame, which only allows certain types of data to be observed.

Like strong and weak sampling, the effects of the different sampling frames may also be implemented via different instantiations of Equation 3. Under category sampling, for example, the sampling domain $\mathcal{X}_S$ represents the restricted category from which exemplars are being sampled, as shown in Figure 11d. Because there are no further restrictions on observations, the denominator in Equation 3 sums to 1 as it does in the general case.. Hence, we can capture category sampling in

[12] For discrete-valued hypothesis $|h|$ represents the number of possible data items that are consistent with the hypothesis. In the continuous case $|h|$ denotes the measure (e.g. "volume") of the hypothesis.

[13] "Strong sampling" has previously been instantiated in a variety of ways, ranging from cases where sample items are selected because they share a property of interest to more explicit forms of pedagogy, where sample items are curated by a teacher to illustrate the extension of a category or property. Although these distinctions can sometimes make a difference (e.g., Shafto et al., 2014), for current purposes all can treated as forms of strong sampling.

[14] The version of the sampling frames model presented here is similar to that outlined by Hayes et al. (2017). Hayes, Banner, Forrester, et al. (2019) developed a more complex variant of the sampling model shown here which included a function learning component that made predictions about property generalization over a continuous dimension, but this added complexity is not relevant for the current aims.

terms of the following likelihood:

$$P(x^{\pm} \mid h, s_{\text{CATEGORY}}) = \begin{cases} \theta P(x) & \text{if } x^{\pm} \in T_h \\ (1-\theta) P(x) & \text{otherwise} \end{cases} \tag{6}$$

This likelihood closely resembles weak sampling (Equation 4); the difference is that because the noise parameter θ may taken on any value in $[0,1]$ no hypothesis is ever entirely falsified. Setting $\theta = 1$, as in the case of noiseless observations, recovers the weak sampling likelihood from Equation 4. Because the likelihood of a single observation is independent of the hypothesis under consideration, observing further instances of the same item, while informative with regard to the noise level θ, is uninformative about whether the property might extend to other items. The fact that all observations happen to belong to the same category has no evidentiary value because the sampling frame s only admits those types of items.

Under property sampling, the sampling frame admits only observations with the target property and the learner must explain the fact that all these observations belong to a single category. Hence, the likelihood under property sampling is given by:

$$P(x^{+} \mid h, s_{\text{PROPERTY}}) = \begin{cases} \dfrac{\theta P(x)}{\theta P(T_h^+) + (1-\theta) P(F_h^+)} & \text{if } x^+ \in T_h^+ \\ \dfrac{(1-\theta) P(x)}{\theta P(T_h^+) + (1-\theta) P(F_h^+)} & \text{otherwise} \end{cases} \tag{7}$$

If we further assume that the observations x are drawn uniformly from the set of possible items $\mathcal{X}$, this simplifies to:

$$P(x^{+} \mid h, s_{\text{PROPERTY}}) = \begin{cases} \dfrac{\theta}{\theta |h| + (1-\theta)(|\mathcal{X}| - |h|)} & \text{if } x^+ \in T_h^+ \\ \dfrac{(1-\theta)}{\theta |h| + (1-\theta)(|\mathcal{X}| - |h|)} & \text{otherwise} \end{cases} \tag{8}$$

which closely resembles strong sampling (Equation 5), except that the possibility of noise in the sampling process means that an observation x may be drawn from the set of true positives (of size $|h|$) or false positives (of size $|\mathcal{X}| - |h|$). This is illustrated in Figure 11c. If the sampling process is noiseless (θ = 1), property sampling and strong sampling are equivalent and the size principle is in full force. Under such conditions, the greatest difference between category and property sampling are expected. When some sampling noise is assumed the bias towards smaller hypotheses consistent with the data is diminished to the point (at $\theta = 0.5$) where property sampling and category sampling are equivalent forms of uninformative sampling.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Ashby, F., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233. https://doi.org/https://doi.org/10.1006/jmps.1995.1021

Bareinboim, E., Tian, J., & Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, *28*(1).

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*(6), 413–423. https://doi.org/10.1037/h0054576

Connor Desai, S., Pilditch, T. D., & Madsen, J. K. (2020). The rational continued influence of misinformation. *Cognition*, *205*, 104453.

Ecker, U., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, (18), 570–578.

Ecker, U., Lewandowsky, S., & Apai, J. (2011). Terrorists brought down the plane! – no, actually it was a technical fault: Processing corrections of emotive information. *Quarterly Journal of Experimental Psychology*, *64*(2), 283–310.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological review*, *57*(2), 94.

Feeney, A. (2018). Forty years of progress on category-based inductive reasoning. *The Routledge international handbook of thinking and reasoning.* (pp. 167–185). Routledge/Taylor & Francis Group.

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. *Psychology of learning and motivation* (pp. 1–55). Elsevier.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis. Third edition.* Chapman & Hall/CRC.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *TiCS*, *20*(11), 818–829.

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* (pp. 488–493). Cognitive Science Society.

Hayes, B. K. (in press). Models of inductive reasoning. *Cambridge handbook of computational cognitive science.* Cambridge University Press.

Hayes, B. K., Banner, S., Forrester, S., & Navarro, D. J. (2019). Selective sampling and inductive inference: Drawing inferences based on observed and missing evidence. *Cognitive Psychology*, *113*.

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K. J., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review*, 1–8.

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 University Press.

Hendrickson, A., Perfors, A., Navarro, D. J., & Ransom, K. J. (2019). Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Cognitive Psychology*, *111*, 80–102.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory.* Appleton-Century.

Johnson, H., & Seifert, C. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimnental Psychcology: Learning, Memory & Cognition*, (20), 1420–1436.

Kemp, C., Navarro, D., & Hayes, B. (under review). Adjustment for selection bias in intuitive reasoning. *Article submitted for publication*.

Lawson, C. A., & Kalish, C. W. (2009). Sample selection and inductive generalization. *Memory & Cognition*, *37*, 596–607.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

Lovibond, P. F., Lee, J. C., & Hayes, B. K. (2020). Stimulus discriminability and induction as independent components of generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(6), 1106.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380.

Medin, D., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316.

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

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

Ransom, K. J., Hendrickson, A., Perfors, A., & Navarro, D. J. (2018). Representational and sampling assumptions drive individual differences in single category generalisation. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *40th Annual CogSci Conference* (pp. 930–935).

Ransom, K. J., Voorspoels, W., Perfors, A., & Navarro, D. J. (2017). A cognitive analysis of deception without lying. *39th Annual CogSci Conference*, 992–997.

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

Razran, G. (1949). Stimulus generalization of conditioned responses. *Psychological Bulletin*, *46*(5), 337–365.

Rogers, T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, *15*, 436–447.

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.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

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

Sloman, S., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*, 223–247. https://doi.org/10.1146/annurev-psych-010814-015135

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*(4), 410–441.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Beh. & Brain Sci.*, *24*(4), 629–640.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Xie, B., Navarro, D. J., & Hayes, B. K. (2020). Adding types, but not tokens, affects property induction. *Cognitive Science*, *44*(9), e12895.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297.