The role of sampling assumptions in generalization with multiple categories

Wai Keen Vong (waikeen.vong@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Andrew T. Hendrickson (drew.hendrickson@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Amy Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Abstract

The extent to which people learning categories generalize on the basis of observed instances should depend in part on their beliefs about how the instances were sampled from the world. Bayesian models of sampling have been successful in predicting the counter-intuitive finding that under certain situations generalization can decrease as more instances of a category are encountered. This has only been shown in tasks were instances are all from the same category, but contrasts with the predictions from most standard models of categorization (such as the Generalized Context Model) that predict when multiple categories exist, people are more likely to generalize to categories that have more instances when distances between categories is controlled. In this current work we show that in both one- and two-category scenarios, people adjust their generalization behavior based on cover story and number of instances. These patterns of generalization at an individual level for both one- and two-category scenarios were well accounted for by a Bayesian model that relies on a mixture of sampling assumptions.

Keywords: sampling assumptions, generalization, category learning

Introduction

The ability to generalize beyond existing data is a basic cognitive capacity that underlies a great deal of human learning, categorization and decision-making (e.g. Shepard, 1987; Tenenbaum & Griffiths, 2001; Nosofsky, 1986). To complete the inductive leap needed for generalization, people must make some kinds of assumptions about how that data was generated or sampled. A learner's sampling assumptions influence the evidentiary value of the data, and thus alters what they should infer based on it.

One natural assumption is that each observed datum has been selected independently and then labeled as a member (or not) of the category or concept to be learned. An example of this is a parent who tries to teach a child what a "ball" is by randomly picking from all of the toys her room then labelling them as balls or not. This possibility, called *weak sampling*, implies that all observations x are equally likely, regardless of what hypothesis h the learner has about the category. Mathematically, this corresponds to the notion that $P(x|h) \propto 1$.

A different type of data generation, known as *strong sampling*, presumes that the data has been selected as a random positive example directly from the category to be learned (Tenenbaum & Griffiths, 2001). A parent who teaches the word "ball" by showing the child many different kinds of balls (but not other toys) is strongly sampling from the category of BALL. The key consequence of strong sampling is that it licenses tighter generalizations with increasing data. This is because each datapoint is more informative about the boundaries of the category. Mathematically, for a hypothesis h that consists of $|h|$ possible category members, the strong sampling model implies that $P(x|h) = 1/|h|$ if the observation x falls within the category, and 0 if it does not.

There is evidence that people are sensitive to sampling assumptions, making tighter generalizations when the data appear to have been strongly sampled (e.g. Xu & Tenenbaum, 2007). However, a number of questions remain unresolved, two of which we address in this paper.

The first, more minor, issue relates to the influence of the cover story. As mentioned, work by Xu and Tenenbaum (2007) suggests that both adults and children change their generalization patterns in response to differences in sampling. This appears to contrast with other work by Navarro, Dry, and Lee (n.d.), which found that although sampling assumptions varied between individual participants (with some assuming strong sampling and others weak), people did not change their behavior according to the cover story they were presented with. One way to resolve the discrepancy between these two studies is to conclude that the data generation process was much more obvious in Xu and Tenenbaum (2007). In that study, participants actually saw instances selected in front of them, whereas in Navarro et al. (n.d.) participants simply read different cover stories. Here we explore whether it is necessary for people to see data being generated in order to change their sampling assumptions, or whether a more salient cover story manipulation would be sufficient.

The second issue we investigate is a more important and more puzzling one. It is generally acknowledged that inductive generalization is very closely linked to categorization. For instance, exemplar models of categorization (e.g., Nosofsky, 1986) are constructed by assuming that the learner uses a simple probabilistic model to generalize from each stored exemplar to a target item. The "narrowness" of the generalizations is a fixed parameter (referred to as the specificity)

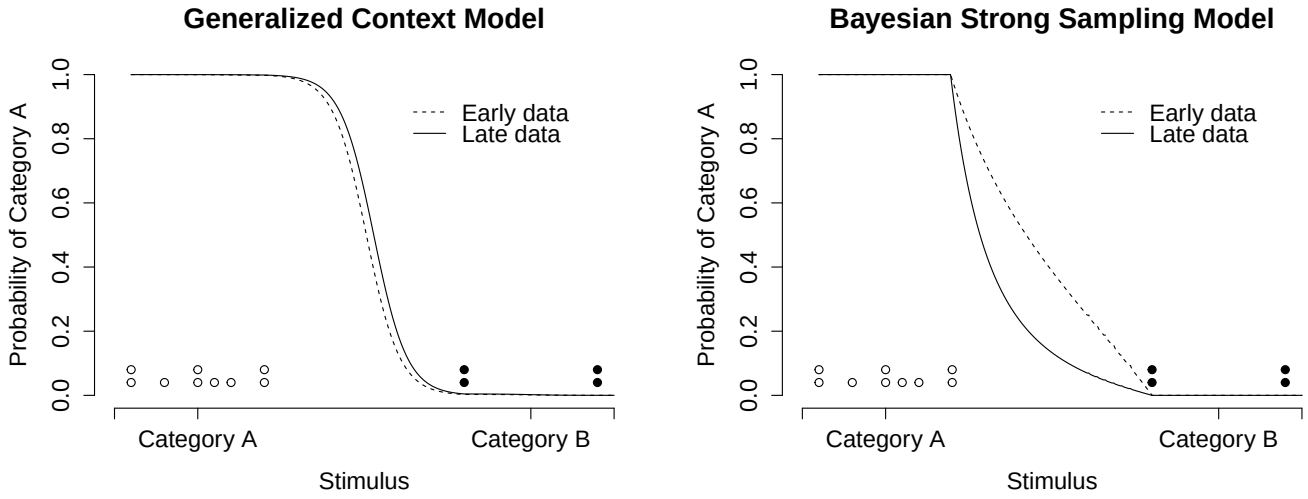

Figure 1: Different predictions made by a standard categorization model (the GCM) and a Bayesian model that incorporates a strong sampling assumption. The GCM, on the left panel, predicts that as the number of instances in a category increases (shown in the figure by the additional points in the bottom row), generalizations should loosen slightly: the solid line corresponding to generalizations based on later additional instances in Category A extends further from Category A. By contrast, the model incorporating strong sampling predicts that generalization based on additional instances will tighten: the solid line in the right panel is much closer to Category A than the dotted line corresponding to earlier, fewer instances.

and does not change as the sample size increases. This is effectively a weak sampling assumption, and it is assumed by both the basic Generalized Context Model (Nosofsky, 1986) and by models such as ALCOVE (Kruschke, 1992) that extend it. These models have proven highly successful at describing human classification behavior, apparently with little need to adapt them to incorporate some version of the strong sampling assumption. If human learners are as sensitive to sampling assumptions as papers such as Xu and Tenenbaum (2007) imply, why has it not been necessary to incorporate such assumptions into existing categorization models?

We can think of at least two possible (not mutually exclusive) explanations for this. The first one is that sampling effects have not been found simply because few studies have gone looking for them. Standard supervised classification designs do not manipulate the sampling assumptions, and it could be argued that the instructions and design of such experiments often imply weak sampling. As such, it is natural to expect that the theories used to explain these experiments would implicitly rely on weak sampling assumptions. A similar suggestion is made by Hsu and Griffiths (2010).

An alternate possibility is that these divergent results arise because of a genuine difference in the nature of the experiments: the number of categories involved. Typical categorization experiments generally involve two categories, with stimuli needing to be classified as belonging to one or the other (Nosofsky, 1986). In contrast, researchers testing sampling assumptions have tended to use tasks in which participants are asked to draw inferences about only a single target category (Xu & Tenenbaum, 2007; Navarro et al., n.d.).

In this paper we test the latter possibility by making use of the fact that strong sampling models make a different prediction from standard categorization models in certain situations. A multiple-category version of a Bayesian generalization model with strong sampling[1] predicts that if we increase the number of instances in Category A without changing the number of exemplars in the other category, items that lie in between the two categories should decrease in their probability of being classified as members of Category A. This is because strong sampling leads to tighter generalization of Category A with more instances, (right panel of Figure 1). Note further that this is the opposite of what one would expect from a standard exemplar model: adding more exemplars to Category A but not to the other category can only increase the summed similarity between Category A exemplars and a target item. As a consequence, items that lie between the two categories should increase in their probability of being classified as members of Category A (left panel of Figure 1).

These distinct predictions motivate our experiment: we present learners with either one-category or two-category generalization problems, presented either in the context of a strong or weak sampling cover story. We predict that when in the context of strong sampling, people will modulate

[1]The two-category Bayesian strong sampling model is a minor modification of the one-category strong sampling model described by Tenenbaum and Griffiths (2001). The only difference between that model and the current one lies in how the hypotheses about the extension of a category (the"consequential regions") are defined. In the two-category model the stimulus space is divided into two mutually exclusive regions, one for each category. As per the original model, category items are assumed to be sampled uniformly at random from the region of that category.

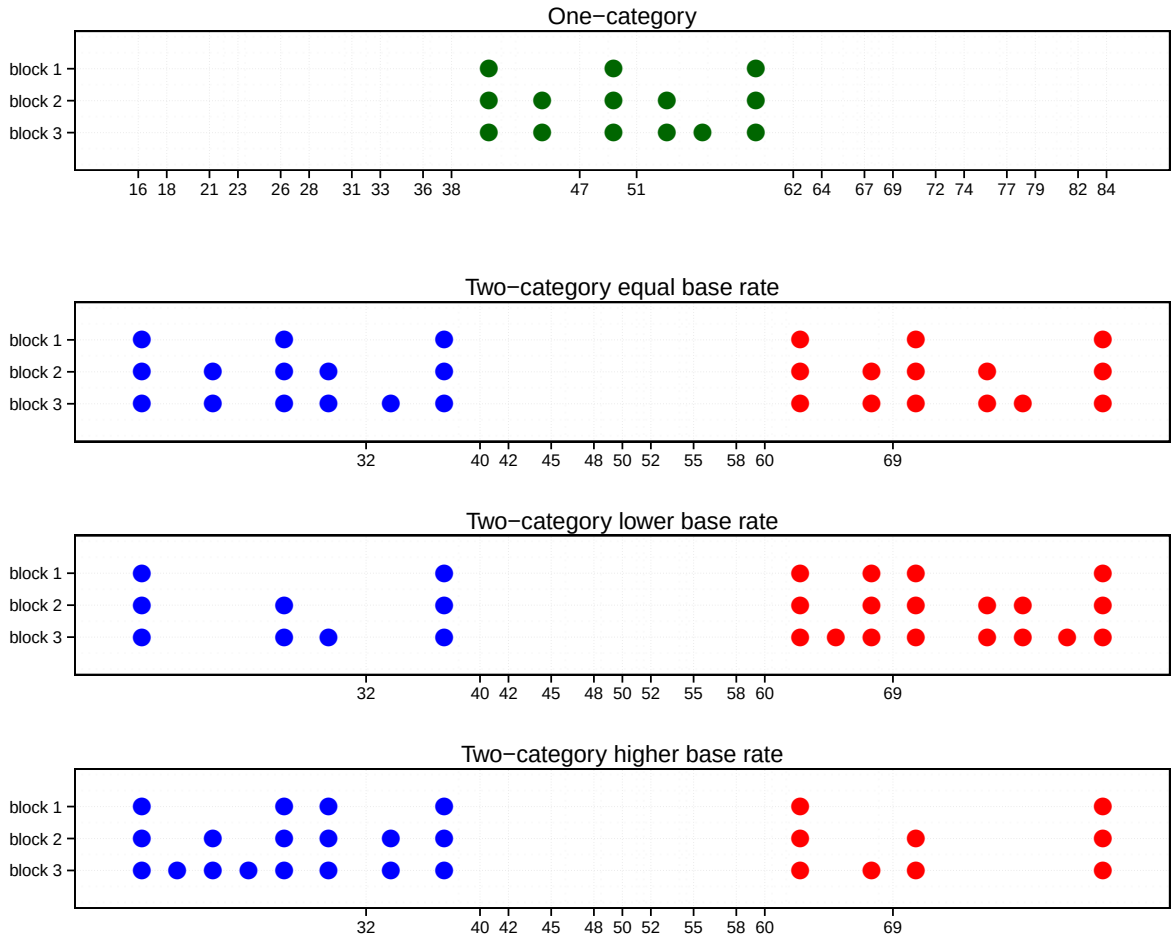

Figure 2: The experimental design. The top panel refers to the three sets of stimuli used across each block in the one-category task. The bottom three panels refer to the three possible sets of stimuli used in the different base rate conditions for the two-category task (results from the bottom two sets are collapsed into one UNEQUAL BASE RATE condition for the purposes of analysis). All participants performed the one-category task as well as one of the three two-category tasks. The ticks at the bottom of each panel show the location of each of the test points for each condition.

their generalization based on cover stories such that they will tighten their generalization of a category label in response to observing additional exemplars in that category that do not extend the category boundary.

Method

Participants Data was collected from 318 participants from Amazon Mechanical Turk. No demographic information was collected so participants remained anonymous. Participants were paid $0.50USD for their participation to complete the task which lasted approximately 15 minutes.

Procedure Each participant performed a one-category and a two-category generalization task in random order following a scenario adapted from Navarro et al. (n.d.). In the one-category task, participants observed instances of temperatures at which one type of bacteria was found alive in food. They were then asked to estimate the probability that the same bacteria would be found alive in the food at other temperatures. In the two-category task, participants observed instances of temperatures where two types of bacteria were found alive in food. They were also told that the two types of bacteria competed for resources, so only one type of bacteria could be found alive in the food at any given temperature. As in the one-category task, participants were asked to estimate the probability that one of the two types of bacteria would be found alive in the food at other temperatures.

The experiment also contained two between-subjects manipulations. The first was a sampling assumption manipulation in which participants were presented with different cover stories in order to influence their beliefs about the sampling process. In the STRONG SAMPLING condition, participants were told that the instances were selected by scientists who had identified a number of temperatures where bacteria were found alive in food. This cover story suggested to the participant that the scientists were only selecting positive examples from the category, consistent with strong sampling. Conversely, in the WEAK SAMPLING condition, participants were told that the instances were the result of an automated pro-

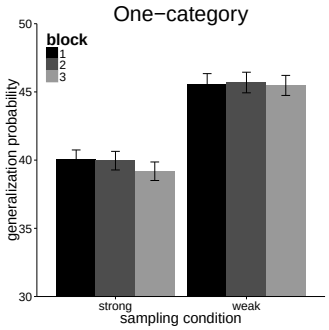

Figure 3: Mean generalization probabilities in the one-category task across sampling condition and block. Generalization is tighter in the STRONG SAMPLING conditions but does not differ by block.

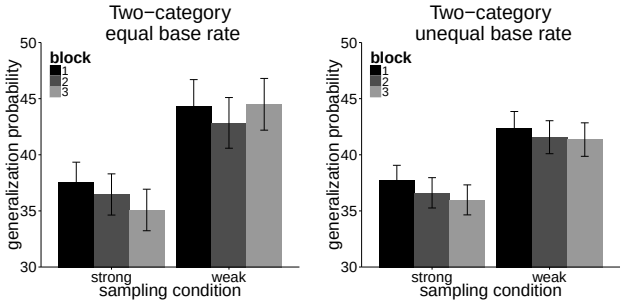

Figure 4: Mean generalization probabilities in the two-category task across sampling condition, base rate condition, and block. Generalization is tighter in the STRONG SAMPLING condition regardless of category base rate.

cess that tested the bacteria at different temperatures. This suggested to the participant that the presented instances were chosen at random from the range of all possible temperatures, consistent with weak sampling. People who were in a given sampling condition received the same (strong or weak) sampling cover story for both the one- and two-category tasks.

The other between-subject experimental manipulation varied the base rate in the two-category generalization task. In the EQUAL BASE RATE condition, the number of instances observed in both categories was the same. There were also two conditions in which one category contained more instances: one in which the left category had more and one in which the right category had more. Because there were no differences between these two conditions, all analyses collapsed them into one UNEQUAL BASE RATE condition. The different conditions are illustrated in Figure 2.

In both the one-category and two-category tasks, the instances were presented across three blocks. In the one-category task, participants initially saw three green dots representing temperatures where bacteria was found alive in the food. They were then asked to estimate (using a slider) the probability that the bacteria would be found alive at each of 22 temperatures in sequence. As a measure of whether participants were performing the task correctly, two of the 22 test trials were located inside the range of observed instances. After making the 22 judgments, participants were then presented with two more instances and asked to make the same judgments again. In the final block, they were presented with one more instance before repeating the 22 estimates again. Overall, each participant made 66 judgments (3 blocks $\times$ 22 queries) in the one-category task.

The procedure in the two-category task was very similar, except that participants were presented with instances representing the temperatures where bacteria from both the left and right categories were found alive (shown as blue and red dots respectively of Figure 2). Participants were asked to estimate the probability that the blue bacteria (the left category) would be found alive in the food at each of 11 temperatures. All of the test points in the two-category task were between the two categories, except for one within the range of instances for each of the categories. As in the one-category case, participants were given additional instances at the beginning of each block and then asked to make judgments at each of the test points. This resulted in a total of 33 judgments (3 blocks $\times$ 11 queries).

Results

Participants who failed to understand the task (based on their performance on the within-category test points) were excluded from the analyses. We reasoned that people who correctly understood the experimental task would have assigned probabilities close to 100% for the test points within the categories. Therefore, participants who assigned a probability of less than 90% on all six test points were removed from that condition. This left 203 participants in the one-category task and 165 participants in the two-category task.

Our first question was whether different cover stories about sampling had an effect on generalization. We examined this by first looking at the raw generalization probability estimates provided by participants. Figures 3 and 4 show the mean generalization probabilities across each condition by block. Consistent with our predictions, in both tasks the mean generalization probability was lower (i.e., participants tightened their generalizations more) in the STRONG SAMPLING condition relative to the WEAK SAMPLING condition ($t(201)$ = -.290, $p < .05$ for the one-category task and $t(163)$ = -3.07, $p < .05$ for the two-category task).

Another way to determine whether the sampling cover story had an effect is to fit individual data using the mixed sampling model from Navarro et al. (n.d.). This model interpolates between weak and strong sampling assumptions, assuming that an observation is strongly sampled with probability θ and weakly sampled with probability $1-\theta$. We can use this to calculate a best-fit θ value for each person, reflecting the extent to which their generalizations were consistent with strong sampling (θ close to 1), weak sampling (θ close to 0), or something in between. Because the mixed sampling

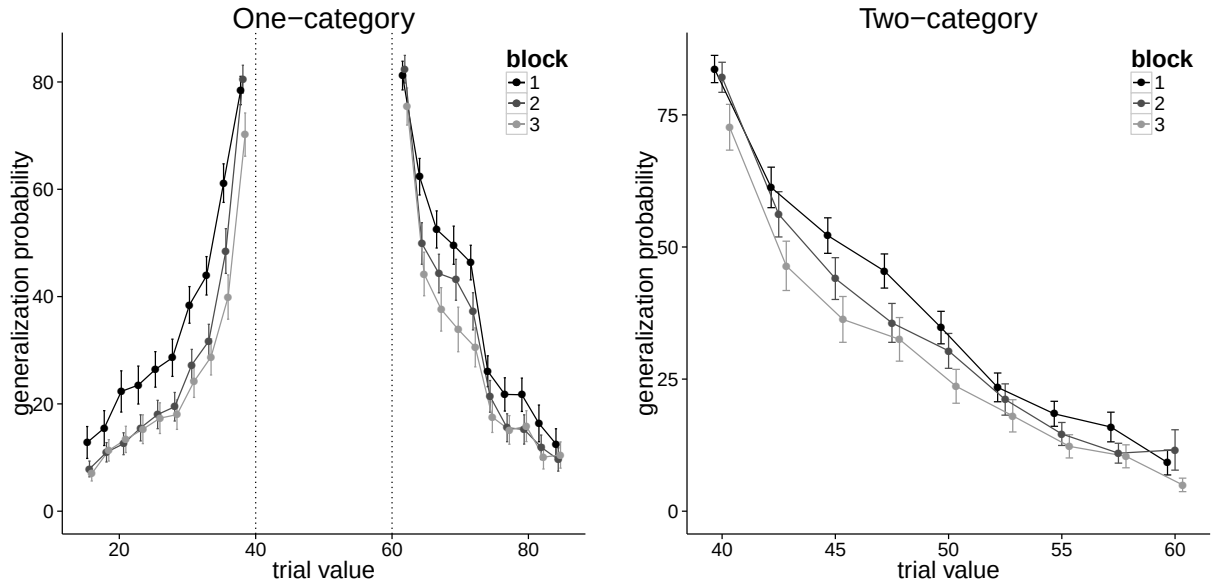

Figure 5: Generalization by block in an additional experiment in which participants were given many more instances in blocks 2 and 3. Generalization probabilities tightened with additional instances, suggesting that earlier lack of tightening was due to conservative updating rather than rejection of the implications of a strong sampling model.

model was originally developed to model generalization responses from a single category, we captured participant responses from the two-category case by treating the left category as the single category whose consequential region is bounded by the leftmost point in the right category. Overall, the model was able to provide a good account for individual responses in both tasks, with a median correlation between the model predictions and participant responses of 0.92 in the one-category task and 0.96 in the two-category task.

As expected, there were significant individual differences in inferred θ values as a function of cover story. Calculating separate beta regressions over condition for each task (which we did because the distribution of θ values deviated from normality) shows that type of sampling condition was a significant predictor of the estimated θ value in both the one-category ($z(3)$ = -4.23, $p < .001$), and the two-category task ($z(3)$ = -4.38, $p < .001$). Overall, these results suggest that people did change their generalizations in response to the cover story, and that the θ parameter in the mixed sampling model is sensitive to that change.

A related prediction was that increasing the number of instances should result in tighter generalization in the STRONG SAMPLING condition. We tested this prediction by comparing generalization probabilities in the first and last (third) block of test trials (shown in Figures 3 and 4). Although there was a significant difference between generalization probability in the first and last blocks in the two-category EQUAL BASE RATES condition (paired-samples t-test, $t(29) = 2.16, p = 0.019$), the differences in the one-category task ($t(116) = 1.08, p = 0.142$) and the two-category UNEQUAL BASE RATES condition ($t(65) = 1.50, p = 0.069$) did not reach significance.[2] Is this because people do not, as predicted by the strong sampling model, tighten their generalizations with additional instances? Or is it simply that people are *conservative*, tightening their generalizations less than such a model would predict?

To investigate this question, we ran an additional experiment involving generalization with 47 participants in the one-category task and 44 participants in the two-category task. The experiment was identical to the STRONG SAMPLING condition of the previous one except that participants were shown many more instances in blocks two and three. As Figure 5 illustrates, when presented with these large amounts of additional instances people in all conditions and tasks tightened their generalizations considerably. Each person's mean generalization probability in the last block was significantly less than their generalization in the first block in both the one-category ($t(46) = 4.53, p < .001$) and the two-category task ($t(43) = 4.07, p < .001$). Within the two-category task, generalizations tightened significantly in both the EQUAL BASE RATES ($t(14) = 2.31, p = 0.018$) and UNEQUAL BASE RATES ($t(28) = 3.27, p = 0.014$) condition.[3] This pattern of tightening with more instances is more consistent with a Bayesian model that includes some proportion of strong sampling than a standard categorization model like the GCM.

[2] As expected, all differences in the WEAK SAMPLING conditions were not significant, with p values ranging from 0.317 to 0.458.

[3] Recall that this condition incorporated the LOWER BASE RATE and HIGHER BASE RATE conditions into one analysis in which both the high-base-rate and low-base-rate left-hand category were combined. Both show significant tightening when analyzed separately as well.

Discussion

This current work clarifies perhaps the most troublesome aspect from Navarro et al. (n.d.): that large individual differences in proportion of strong and weak sampling assumptions were observed but people did not seem to be sensitive to the sampling type suggested by the cover story. By directly referencing in the cover story how samples were being generated: either by direct selection for strong sampling or random occurrence for weak sampling, we find reliable differences in generalization between the two cover stories in the one-category condition. This pattern of results is accounted for naturally by a Bayesian model using a mixture of strong and weak sampling assumptions (Navarro et al., n.d.) and is consistent with standard categorization models such as the GCM (Nosofsky, 1986) that rely on differences in the specificity parameter between cover story conditions. We believe the sampling assumption model account is slightly more parsimonious because it *a priori* predicts that the weak sampling condition will show wider generalization gradients than the strong sampling condition, rather than relying on a freely varying model parameter.

Interestingly, the difference between strong and weak cover stories is found not only in the one-category but also the two-category scenario. That this pattern exists not only when only positive examples of a single category are observed but also when more than one category is observed, suggests that beliefs about sampling processes influence behavior even in situations more traditionally thought of as category learning. As in the one-category scenario, a model Navarro et al. (n.d.) without category learning processes and relying only on different mixtures of sampling assumptions is able to account for the behavioral results with a high degree of accuracy.

The presence of significant gradient tightening only at large changes in the number of instances suggests some additional process is mediating the effect of gradient tightening predicted by the Bayesian model that incorporates a mixture of strong and weak sampling. One possibility for such a mediating process would be conservatism (Phillips & Edwards, 1966), some reluctance to update beliefs about the boundary of each category as much as is suggested by a rational model that includes strong sampling. This conservatism may be due to assumptions that learners might be making about other possible sampling processes including noisy instance generation, noisy labelling, or could be the result of cognitive processes that do not weigh each instance equally as the Bayesian model does (Navon, 1978).

In summary, in both the one- and two-category scenarios, people had different patterns of generalization from known instances to new instances based on a cover story that suggested strong or weak sampling was generating the instances they saw. Additionally, the degree of generalization decreased as many more instances were shown from the target category, more than predicted by standard models of categorization like the GCM but less than predicted by a Bayesian model that mixed strong and weak sampling. Patterns of generalization at an individual level for both one- and two-category scenarios were well accounted for by this Bayesian model, suggesting people are sensitive to the sampling assumptions that are generating the instances they see during categorization.

Acknowledgments

DJN, AP, and ATH were all supported by ARC grant DP110104949. In addition, DJN received salary support from ARC grant FT110100431 and AP received salary support from ARC grant DE120102378.

References

Hsu, A., & Griffiths, T. (2010). Effects of generative and discriminative learning on use of category variability. In *32nd Annual Conference of the Cognitive Science Society.*

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22.

Navarro, D. J., Dry, M. J., & Lee, M. D. (n.d.). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*.

Navon, D. (1978). The importance of being conservative: Some reflections on human bayesian behaviour. *British Journal of Mathematical and Statistical Psychology*, *31*(1), 33–48.

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

Phillips, L., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology: General*, *72*(3), 346.

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–641.

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