Category number and frequency shape sampling assumptions

Andrew T. Hendrickson
Cognitive Science & Artificial Intelligence
Tilburg University

Amy Perfors
School of Psychology
University of Adelaide

Danielle J. Navarro
School of Psychology
University of New South Wales

Keith Ransom
School of Psychology
University of Adelaide

Abstract

Categorization and generalization are fundamentally related inference problems. Yet leading computational models of categorization (as exemplified by, e.g., Nosofsky, 1986) and generalization (as exemplified by, e.g., Tenenbaum & Griffiths, 2001) make qualitatively different predictions about how inference should change as a function of the number of items. Assuming all else is equal, categorization models predict that increasing the number of items in a category increases the chance of assigning a new item to that category; generalization models predict a decrease, or a tightening with additional exemplars (this is known as the size principle). This paper investigates this discrepancy, showing that people do indeed perform qualitatively differently in categorization and generalization tasks even when all superficial elements of the task are kept constant. When there are two categories, people behave in accordance with models of categorization and increase the chance of assigning more distant items to that category. When there is one category, people behave in accordance with models of generalization and tighten their inferences. We show that neither model can naturally account for the pattern of behavior across *both* categorization and generalization tasks, and discuss the implications of these results for understanding category learning.
Keywords: Categorisation, Generalisation, Similarity, Inference, Cognitive modeling.

Introduction

Categorization and generalization are two fundamental and deeply related inductive problems. Categorization requires people to determine which out of a set of labels should be applied to a novel object, whereas generalization problems ask the learner to determine whether a particular label should be applied to a novel object. The surface differences between the two tasks appear to be almost negligible and in one sense are purely a matter of framing. In a typical supervised categorization task a participant is usually asked a question like *"is this item a Dax or a Wug?"*. A generalization task, on the other hand, is more likely to be framed as a question about the extension of a single target category: the query is more likely to be *"is this item a Dax?"*. In a category learning task where every object belongs to exactly one of two categories it is possible to reduce both problems to the same inductive problem in which the goal is to determine whether the novel object is a Dax or a non-Dax.

Viewed from this perspective, one might expect categorization and generalization to be essentially identical. Both require the learner to make inferences about the extensions of categories, both predict people's behavior on the basis of psychological theories about how categories are represented, and both depend on the learner having formed some representation of the categories on the basis of a set of exemplars. Accordingly, one would expect that theories of categorization and theories of generalization should agree with each other, at least qualitatively, when describing the inferences people make. In this paper we investigate a surprising and robust disagreement between these two different inference problems and show how this difference is mirrored in existing theoretical accounts. Specifically, we show that increasing the *sample size* has qualitatively different effects on human inductive inferences in categorization and generalization.

To illustrate why we might predict the effect of sample size to differ across tasks, we consider each task separately. In a Dax-or-Wug **categorization** problem, increasing the number of Dax observations (holding other factors constant) pushes the category boundary away from the observed Dax exemplars. Theoretical models of categorization capture this frequency effect in a natural fashion. For example, the Generalized Context Model (GCM; Nosofsky, 1986) is an exemplar model of categorization that computes a response strength for the Dax category by summing the similarities between the novel object and every previous Dax exemplar. Accordingly, adding more Dax observations without adding any Wug exemplars will increase the strength of the Dax category, especially for items near the Dax observations or whose category label is ambiguous. An item that was previously equally likely to be classified as Dax or Wug will now appear more Dax-like because additional Dax exemplars have been added. To put it another way, the GCM predicts a base rate effect in which the point of subjective equivalence (where the response strengths for the two categories are equal) is pushed away from the Daxes and towards the Wugs.

Now consider a **generalization** problem in which a learner is shown several Daxes and asked to determine whether a novel item is also a Dax. What happens to people's generalizations as we increase the number of Daxes? By analogy to the categorization problem one might suppose that more examples of Daxes would encourage people to generalize more broadly. However, formal models of generalization – e.g., the Bayesian approach taken by Tenenbaum and Griffiths (2001) – predict precisely the opposite. As the learner encounters more Daxes they become more confident that the empirically observed variation in Daxes is entirely representative of the full range. When only a few Daxes have been seen, it is quite plausible to believe that a novel object is also a Dax, even if it is somewhat dissimilar to the previously encountered item. Observing one two-inch-tall Dax and one three-inch-tall does not rule out the possibility that Daxes can be four inches; but if the learner has seen 100 Daxes, none larger than three inches, the odds that Daxes can be four inches become much lower. If Daxes that large were possible one should have encountered them by now, and as a consequence the learner shows very little generalization to new items that differ significantly in size.

Despite the apparent inconsistency, both effects have found substantial empirical justification. Increasing frequency consistently produces tightening of generalizations across a variety of experimental frameworks and contexts (Tenenbaum, 1999, 2000; Sanjana & Tenenbaum, 2003; Xu & Tenenbaum, 2007b, 2007a; Frank & Tenenbaum, 2011; Lewis & Frank, in press; Navarro & Perfors, 2010; Navarro, Dry, & Lee, 2012; Vong, Hendrickson, Perfors, & Navarro, 2013; Hsu & Griffiths, 2016). Increasing frequency in categorization has produced a variety of expansion effects including expanding category generalization (Nosofsky, 1991, 1988b; Harris, Murphy, & Rehder, 2008) and typicality judgments (Vandierendonck, 1988; Williams & Durso, 1986) when a single item within a category is repeated, increased stability and generalization (Donald, Joseph, Don, David, & Steven, 1973; Homa & Vosburgh, 1976; Breen & Schvaneveldt, 1986; Homa, Burruel, & Field, 1987; Homa, Dunbar, & Nohre, 1991) as the frequency of all categories increases, as well as the expansion of category membership predictions (Barsalou, 1985), category size estimates (Beyth-Marom & Fischhoff, 1977), trait acceptance (Boseovski & Lee, 2006), and relative similarity (Polk, Behensky, Gonzalez, & Smith, 2002). This is somewhat surprising: the implication is that the same manipulation (increasing sample size of the Dax category) causes the Dax category to expand when items from two categories are shown and the task is framed as a Dax-or-Wug problem, but causes it to tighten when items from one category are shown and the task is recast as Dax-or-not-Dax. It becomes more surprising when one realizes – as we demonstrate later – that neither model predicts this reversal. The original GCM predicts expansion in the categorization task, and the natural adaptation of the GCM to a generalization task continues to predict expansion. Similarly, the Bayesian generalization model predicts narrowing in the original problem and continues to do so when applied in a Dax-or-Wug style categorization task.

Given how puzzling the inconsistency appears, one might suppose that it could be resolved by showing that one of the two phenomena is an experimental artifact. Perhaps the difference can be attributed to different choices of stimuli, different choices of dependent measure, or different kinds of presentation. Our goal in this paper is to present experiments that eliminate these differences and assess categorization and generalization experiments using a common experimental paradigm. In doing so, we provide clear empirical evidence that people do in fact treat these problems in different ways.

The structure of our paper is as follows. We begin with a more careful discussion of the theoretical issue, showing how the inconsistency between the two modeling approaches arises because of a fundamental difference in how they conceptualize the inference problem and is not due to to superficial modeling choices like parameter settings. We then present two experiments that show that the effect of sample size is indeed different in the categorization task than in the generalization task even when using common stimulus sets and response measures. We argue that the difference arises because there is a genuine difference between the two problems: figuring out how to generalize from one category is a qualitatively different kind of thing than figuring out how to categorize between two. Finally, in a third experiment, we show that these frequency effects are modulated by instructional manipulations that influence the prior beliefs about how items are sampled, suggesting a common cognitive mechanism.

Before continuing, given that there is some ambiguity about the meaning of terms like categorization and generalization, it is important to be precise about how we are using the terms. Throughout the paper we use *categorization* to refer to the inference problem in which items from more than one category are encountered during training, while *generalization* denotes the problem in which learners must make judgments when only seeing items from one category. Another important distinction we make is about the nature of the dependent measure used to predict human performance. We use the term *forced choice task* whenever the dependent measure is constructed from a forced choice decision (either Wug-or-Dax or Wug-or-Not-Wug). Conversely, a *probability judgment task* refers to situations where the dependent measure is a probability judgment of membership in a single category.

Models of generalization and categorization

We begin by systematically evaluating two specific models of categorization and generalization. Do they genuinely produce these different effects, and if so why?

On the categorization side, we focus on the generalized context model (GCM) of Nosofsky (1986). We choose this model because it is the archetypical model within the categorization literature. It has been used to account for a wide range of phenomena in categorization including including item and category frequency effects (Nosofsky, 1988b, 1991), typicality (Nosofsky, 1988a) and distortion (Zaki & Nosofsky, 2007) in category inference, and the reaction times of judgments (Nosofsky & Palmeri, 1997). It is also representative of other categorization models in terms of the qualitative behavior in question. As we discuss later, a large range of categorization models – including prototype and prototype-hybrid models, decision-boundary models, knowledge-partitioning models, and Bayesian category-learning models – match the prediction of the GCM that a learner should be more likely to assign a novel exemplar to a category when there are more items in the category. For simplicity, rather than analyze all of these models, we center our attention on the GCM.

On the generalization side we focus on the Bayesian generalization model of Tenenbaum and Griffiths (2001). We choose this model because it was the first computational model in the generalization literature to propose a mechanism (the size principle) to account for generalization gradient tightening. This effect has been observed in a number of experimental contexts including concept learning (Tenenbaum, 2000, 1999; Sanjana & Tenenbaum, 2003; Navarro & Perfors, 2010), language learning (Xu & Tenenbaum, 2007b, 2007a; Hsu & Griffiths, 2016; Lewis & Frank, in press; Frank & Tenenbaum, 2011), and category generalization (Navarro et al., 2012; Vong et al., 2013). As with the GCM and models of categorization, rather than implement the variety of extensions and related models, we choose for simplicity to focus on a single model.

There is another reason to discuss the GCM and the Bayesian generalization model in particular: these models are theoretically related to each other. The shared common core of both models is Shepard's (1987) theory of generalization, and both models extend this theory in different ways. Shepard's analysis argues that the probability of generalizing from a single observed entity decreases exponentially as a function of distance in an appropriately constructed psychological space. The GCM extends Shepard's analysis by applying an exponential generalization gradient to all exemplars, and using the summed generalization strengths to guide categorization decisions. The Bayesian generalization model is also related to Shepard's model, retaining its central constructs but reformulating generalization as a Bayesian inference problem. Like Shepard, it assumes that there exists some true extension of the unknown category – the "consequential region" – and the learner's goal when generalizing from a set of exemplars is to estimate the probability that a novel item falls within the consequential region of the psychological space.

In all three theories (Shepard, Nosofsky, Tenenbaum & Griffiths) there is a critical link between inferring category memberships and making generalizations. However, while the GCM and the Bayesian generalization model can both be viewed as extensions to Shepard's law of generalization, they are not equivalent. The GCM is primarily a model of categorization that can be adapted to generalization problems, whereas Tenenbaum and Griffiths provided a Bayesian model of generalization that is extensible to categorization. At it turns out, this is the critical difference that produces the inconsistency.

GCM: An exemplar model for categorization

The central theoretical idea in the generalized context model is that the learner stores copies of all observed exemplars and generalizes from them separately. Following the approach of Shepard (1987), the GCM assumes that stimuli are represented as points in a psychological space and the similarity between stimuli decreases exponentially with distance. Suppose the learner has observed a set of N exemplars $\boldsymbol{x} = (x_1, \ldots, x_N)$ and category labels $\boldsymbol{l} = (l_1, \ldots, l_N)$, where the label l_i for the i-th item belongs to a set of K possible category labels. The learner then encounters a new item y and must decide which of the K categories it belongs to. If we let $d(x, y)$ denote the distance between two items in psychological space, then the GCM uses Shepard's exponential generalization gradient as a method to define the similarity $s(x, y)$, as follows:

$$s(x, y) = \exp(-\lambda d(x, y)). \tag{1}$$

In this expression λ denotes the *specificity* parameter that describes the steepness of the generalization gradient.[1] The GCM uses this similarity function to determine the response strength $\eta(y, c)$ for a particular category c when the learner is presented with a test item y. Following the core principle of exemplar models (Medin & Schaffer, 1978), the GCM assumes that the learner separately assesses the similarity between the test item, and these similarities combine additively. This gives rise to the following sum-similarity rule:

[1]When items vary on multiple stimulus dimensions the GCM applies dimension weights to incorporate differences in attention or feature salience, and applies an appropriate Minkowski metric (Euclidean distance for integral dimensions, city block distance for separable ones) to compute distance. However, in all our experiments stimuli vary only on a single dimension and distances can be computed without additional parameterization.

$$\eta(y,c) = \sum_{x_i | l_i = c} s(y, x_i) \tag{2}$$

where the sum is taken across all exemplars that belong to category c.[2] Following (Luce, 1963) the probability of assigning the test item y to category c is assumed to be proportional to the response strength, giving:

$$P(l_y = c) = \frac{\eta(y,c)}{\sum_{c'=1}^{K} \eta(y,c')} \tag{3}$$

where the sum in the denominator is taken over all categories present in the task. Notwithstanding the subtleties associated with different dependent measures[3], we take it that Equation 3 describes the GCM predictions: in a judgment task we assume that people directly report the value of $P(l_y = c)$ plus some response noise, and in a forced choice task we assume that $P(l_y = c)$ describes the probability of selecting category c. Neither assumption is likely to be correct, of course, but for the purposes of our paper it suffices to note that there is no reason to think this is important for the purposes of considering the effect of sample size.

The GCM makes a very clear prediction about the effect of relative category frequency in categorization. As Figure 1 illustrates, increasing the size of one category relative to another causes a base rate effect: all else being equal, the fact that category A has more exemplars than category B makes it more likely that the model assigns the label A to any particular test item. The net effect is that the category boundary is "pushed" away from the category A exemplars and towards the category B exemplars. Note that although the specificity parameter λ can influence the shape of the curves, the qualitative prediction is invariant to the value of λ. As long as the range of stimulus space spanned by the observed exemplars is kept constant, the GCM will never predict an effect in the opposite direction.

Moreover, it should be noted that this prediction is not specific to the GCM or exemplar models. Rather, it is grossly typical of categorization models generally. For instance, prototype-hybrid models shift the prior beliefs of the category of new items to match the empirical category frequency or create new prototypes within the more frequent category (Anderson, 1990, 1991; Sanborn, Griffiths, & Navarro, 2010, 2006; Love, Medin, & Gureckis, 2004). Decision-boundary models (Ashby & Gott, 1988; Ashby & Perrin, 1988; Ashby & Townsend, 1986; Ashby & Maddox, 1993) and knowledge partitioning models (Lewandowsky & Kirsner, 2000; Lewandowsky, Kalish, & Griffiths, 2000; Yang & Lewandowsky, 2004) either do not change with increased frequency, or shift their representations to increase the portion of the stimulus space that corresponds to the more frequent category. Even a Bayesian category learning model based on rule-inference will increase

[2]More precisely, the sum in Equation 2 should be taken across all distinct presentations of each stimulus, so that if stimuli have different presentation frequencies they can be weighted differently. The distinction is irrelevant for our studies as we ensure that all items have the same presentation frequency.

[3]For instance, in some applications of the GCM to forced choice tasks, a response scaling parameter γ is added, in which case the response strength is given by $\eta(y,c)^\gamma$. Response scaling has been the focus of some discussion in the literature (e.g. Smith & Minda, 2002, 1998; Navarro, 2007; Myung, Pitt, & Navarro, 2007; Nosofsky & Zaki, 2002), but for our current purposes it plays no meaningful role and can safely be omitted.

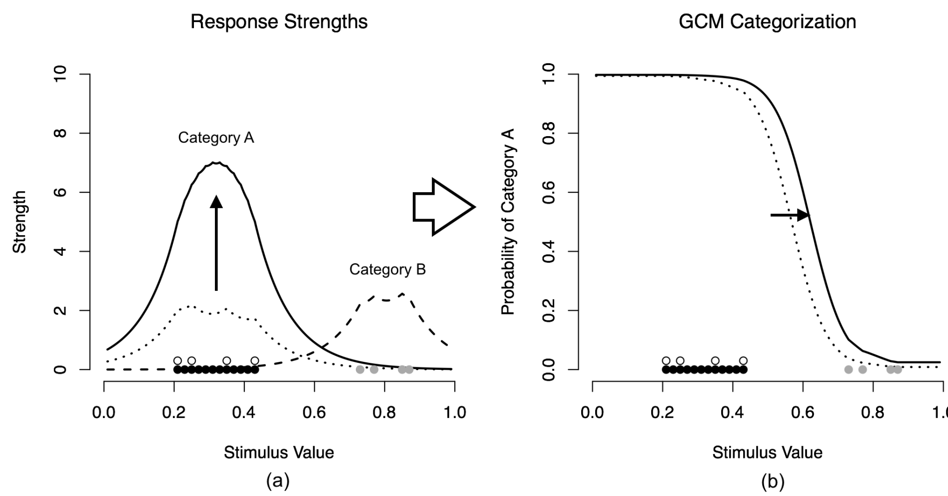

Figure 1. The effect of sample size on the GCM in a categorization task. The left side of panel a shows how the response strength increases when the number of exemplars of a target category (category A) is increased from 4 (white markers, dotted line) to 12 (black markers, solid line). The number of exemplars of category B (grey markers) is held constant at $n = 4$, and so the response strength (dashed line) remains unchanged. The effect on categorization is shown in panel b: because the response strength for category A increases relative to category B, the category boundary is pushed to the right.

the complexity of the category rule with more items, leading to increased generalization for new items (Goodman, Tenenbaum, Feldman, & Griffiths, 2008).

Applying the GCM to a generalization problem

The GCM in the form described by Nosofsky (1986) is primarily a model for how people choose *which* category to assign a novel item. Thus, although it is related to Shepard's theory of generalization, it is not itself a model for stimulus generalization. It provides answers the "Dax or Wug?" question, not the "is this a Dax?" question. Nevertheless, it is not difficult to extend the GCM to make a prediction of this form, and indeed this extension was proposed by Nosofsky (1991) not long after the GCM was originally developed. The intent at the time was to adapt the GCM to serve as a model of recognition memory. Recognition memory experiments have much the same structure as a generalization task: people are shown items that belong to a single list (the target category) and asked whether a test item was found on the study list. The two problems are not perfectly equivalent in that the recognition memory task asks for an identification decision rather than an inductive generalization, but there is mounting evidence (Nosofsky, 1991; Hawkins, Hayes, & Heit, 2016; Nosofsky, 2016; Nosofsky, Cox, Cao, & Shiffrin, 2014) that the underlying processes between recognition memory and induction may be the same. With that in mind, we suggest that the Nosofsky (1991) model represents the natural way to adapt the original GCM to a generalization problem (though see Nosofsky, Little, Donkin, and Fific (2011) where this threshold depends on a stimulus property, an adaptation we explore in the discussion).

The model works in the following way. If the learner has observed multiple stimuli that all belong to the target category c, a response strength for that category $\eta(y, c)$ is computed using Equation 2 above, with no differences from the categorization context. However, because the generalization problem does not provide a contrast category, the generalization probability is computed by comparing $\eta(y, c)$ to a threshold ϕ:

$$P(y \in c) = \frac{\eta(y, c)}{\eta(y, c) + \phi} \tag{4}$$

On its face, this seems a sensible – even natural – adaptation of the GCM. It retains all the core theoretical constructs that the original GCM used to solve a categorization problem, and the only novel entity is a threshold parameter that has proven successful in adapting GCM in very closely related designs. What does it predict?

The behavior of the GCM in a generalization design is plotted in Figure 2, and unsurprisingly the effect is closely analogous to the pattern shown in Figure 1. If we increase the number of exemplars of the target category without changing the range they span, the response strength rises (panel a). As a consequence, if the threshold parameter ϕ is unchanged, the overall effect is to push the generalization gradient (panel b) outwards, away from the observed exemplars. It should be noted that if ϕ is allowed to change as a function of sample size the GCM makes no prediction at all, insofar it is always possible to choose values of ϕ that move the generalization gradient in any direction. Given this we suggest that – absent a substantive theory about how the threshold parameter ϕ should change – the *default* prediction of the GCM in a generalization design should be expansion rather than narrowing. Later in the paper we discuss a more principled way of adapting GCM to inductive generalization tasks based on the principles from Nosofsky et al. (2011).

The Bayesian model for generalization

A different perspective on inductive generalization is suggested by the Bayesian model of Tenenbaum and Griffiths (2001). Like the GCM, this approach can also be viewed as an extension of the work of Shepard (1987) on stimulus generalization. Unlike the GCM, it is a more direct extension. This framework assumes that a learner is given a set of N items $\boldsymbol{x} = (x_1, \ldots, x_N)$ that all belong in the same category. The learner's goal is to infer whether that category generalizes to include a novel item y. The learner accomplishes this by constructs a hypothesis space $\mathcal{H}$ consisting of a set of "consequential regions" of the psychological space, where each possible region is a candidate for the consequential region that defines the category. They assign some prior degree of belief $P(h)$ to each such hypothesis, which is updated to a posterior distribution via Bayes' rule:

$$P(h|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|h)P(h)}{\sum_{h' \in \mathcal{H}} P(\boldsymbol{x}|h')P(h')} \tag{5}$$

where $P(\boldsymbol{x}|h)$ is describes the likelihood that the learner would have observed the items $\boldsymbol{x}$ if h were indeed the true extension of the category. The generalization probability is then constructed by summing the posterior probabilities of those hypotheses that contain the novel item y:

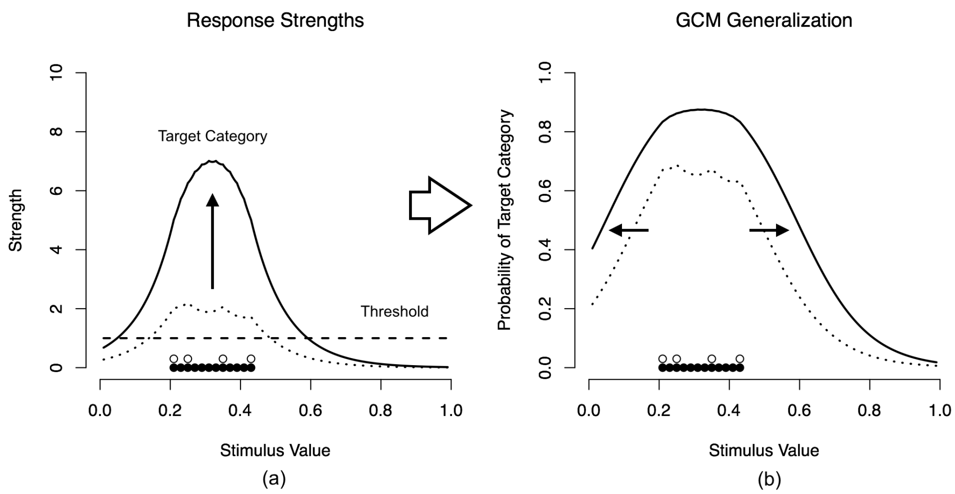

Figure 2. The effect of sample size on the GCM in a generalization task. The response strengths for the target category are shown in the left panel, and are identical to those depicted for category A in Figure 1. However, because the generalization task does not present exemplars from a contrast category, the response strength is compared against a fixed threshold. The effect on generalization is shown in the right panel: increasing the sample size pushes the generalization gradient outwards, away from the observed exemplars.

$$P(y \in c) = \sum_{h|y \in h} P(h|\boldsymbol{x}) \tag{6}$$

As Tenenbaum and Griffiths (2001) point out, this framing of the generalization problem includes Shepard's model as a special case, and has many connections with theories of similarity. Inference in the Bayesian generalization model is driven by the likelihood function $P(\boldsymbol{x}|h)$, which provides the mechanism for belief revision in the model. The model assumes that items are sampled in a conditionally independent manner, which allows the probability of several items $\boldsymbol{x}$ to be expressed as the product of their individual probabilities:

$$P(\boldsymbol{x}|h) = \prod_i P(x_i|h) \tag{7}$$

One of the major departures from Shepard's original model lies in the way that this likelihood function is constructed. Shepard argued that a typical learning scenario was one in which nature selects the item x independently of the consequential region h, yielding what has been termed a "weak sampling" model (Tenenbaum, 1999). This weak sampling model does not change the relative belief assigned to any hypothesis based on the number of items that belong in the category. However, Shepard's weak sampling model is untenable in a situation when *many* items are all constrained to belong to the same category. Given this, Tenenbaum and Griffiths (2001) make the simplest possible alteration to Shepard's assumption of independence: their "strong sampling" model assumes that items are sampled such that they are constrained to belong to the relevant category but are otherwise chosen uniformly at random. This slight change introduces a dependency in the likelihood function on the structure of the hypothesis. Specifically, if the hypothesized category has size $|h|$, then this means that:

$$P(x_i|h) = \begin{cases} 1/|h| & \text{if } x_i \in h \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

In one sense the shift from weak sampling to strong sampling is trivial: it merely incorporates a sensible constraint imposed by the fact that the generalization problem now can incorporate multiple items from the same category. However, it entails a very non-trivial consequence known as the *size principle*: as the sample size N increases, the learner will prefer smaller hypotheses over large ones (Tenenbaum, 1999). To see why this holds, consider the relative degree of belief that the learner has in two hypotheses h_1 and h_2 after observing N items. Assuming that both hypotheses are consistent with all the observations, then:

$$\begin{aligned} \frac{P(h_1|\boldsymbol{x})}{P(h_2|\boldsymbol{x})} &= \frac{P(h_1)}{P(h_2)} \times \prod_{i=1}^{N} \frac{P(x_i|h_1)}{P(x_i|h_2)} \\ &= \frac{P(h_1)}{P(h_2)} \times \prod_{i=1}^{N} \frac{1/|h_1|}{1/|h_2|} \\ &= \frac{P(h_1)}{P(h_2)} \times \left(\frac{|h_2|}{|h_1|}\right)^N \end{aligned} \tag{9}$$

Unlike a learner that assumes weak sampling and has no preference based on hypothesis size, if the two hypotheses are different sizes, the learner who assumes strong sampling will come to prefer the smaller one. Moreover, the extent of this preference grows exponentially larger as the sample size N increases. The net result of increasing the sample size is that the learner shifts belief from large hypotheses to small ones and the generalization gradient tightens. This effect is depicted visually in Figure 3. The tightening of generalizations with increased samples is a direct result of assuming the strong sampling model, a fundamental characteristic of this model. While there are a number of ways in which the model could be adapted by modifying the hypothesis space, the prior, or the sampling model, by default the Bayesian generalization model generates the opposite prediction about the effect of sample size on generalization than the GCM.

Applying the Bayesian generalization model to categorization

Though originally conceptualized in terms of finding the consequential regions for one category, the Bayesian generalization model is trivially extensible to two categories, by simply calculating the generalization separately for each. If the positive examples for each category are modeled as an independent Bayesian generalization process, the posterior estimates can be combined without regard to the frequency of each category by adapting Equation 3

$$P(l_y = c) = \frac{P(y \in c)}{\sum_{c'} P(y \in c')} \tag{10}$$

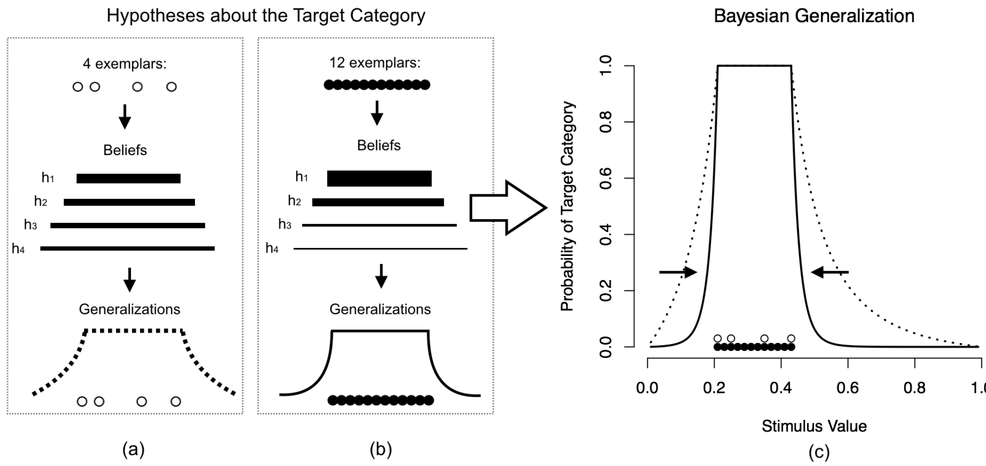

Figure 3. The effect of sample size on Bayesian generalization model. In the case of observing 4 exemplars (top of panel a) drawn from a target category, it is somewhat plausible to think that the true extension of the category might be much broader than the range spanned by those items. Illustrating this idea, the middle of panel a shows the relative degree belief in four hypotheses (horizontal lines), where the width of each line reflects the amount of belief in that hypothesis. By averaging over their beliefs about the hypothesis (bottom of panel a) the learner obtains a broad generalization gradient. In the case of 12 exemplars that span the same consequential region (as in panel b) this evidence very strongly favors smaller hypotheses, producing more belief in smaller hypotheses and thus narrower generalization gradients. Panel c show the generalizations made by a Bayesian learner who considers all possible intervals defined over a finite range for the 4 and 12 exemplar cases (see Navarro, et. al. (2012) for analytic expressions).

where $P(y \in c')$ describes the generalization gradient inferred for category c' using Equation 6. As the number of items in category c increases and the generalization gradients $P(y \in c)$ tightens, the category boundary will also tighten toward c. Perhaps a more natural extension of the strong sampling model for a categorization task is one where the samples are modeled as a draw from all categories. That is, assume that strong sampling occurs within a category, but for each item to sample, first a category is selected to sample from. This choice is done in proportion to the base rates of the categories. This yields an alternative form for Equation 10

$$P(l_y = c) = \frac{P(c)P(y \in c)}{\sum_{c'} P(c')P(y \in c')} \tag{11}$$

where the prior probability of category c is proportional to the number of times it has been observed n_c.

The predictions about sample size that emerge from this model are somewhat less obvious than the previous one because there are two different mechanisms involved: as described before the generalization gradients $P(y \in c)$ tend to tighten as the sample size increases (causing category boundaries to *tighten*). However, this effect is somewhat offset

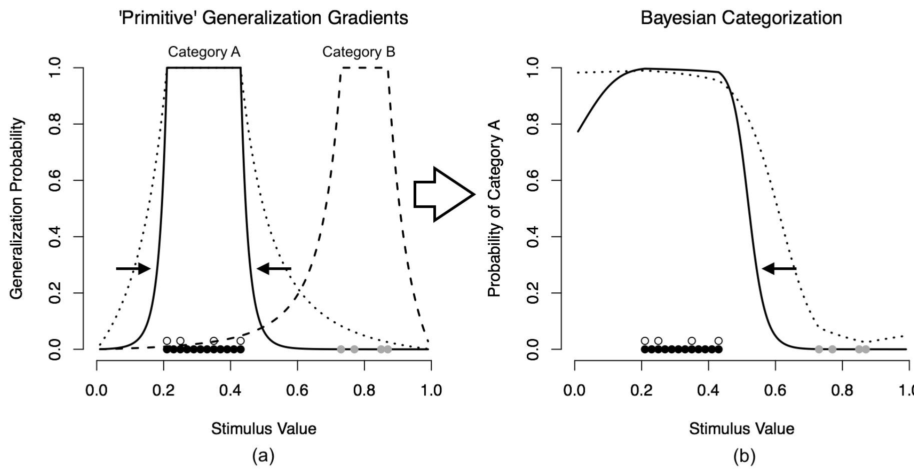

Figure 4. The effect of sample size on Bayesian categorization. If the learner learns about the extension of each category independently, applying the same strong sampling model to explain how observations are selected from each category, the resulting "primitive" generalization functions are shown in panel a. Even when combined with a mechanism for learning the different base rates of the two categories, the overall result when this model is applied to categorization is to shift the category boundary closer to category A when the sample size is increased.

by the fact that the prior probability of the label $P(c)$ *increases* with sample size (causing boundaries to *expand*). Nevertheless, as Figure 4 shows, the first effect dominates the second and the overall effect of increasing sample size is tightening for the category A boundary. Extending the Bayesian model of generalization to a categorization problem produces a model that makes the exact opposite prediction about sample size as the GCM.

How should we interpret the discrepancy?

The inconsistency between these models is striking enough on its own, but becomes even more surprising when one recognizes that the GCM is known to make good predictions about how frequency information is used in categorization (Nosofsky, 1991, 1988b; Navarro & Kemp, under review), and that the Bayesian strong sampling model makes accurate predictions in inductive generalization tasks (Navarro & Perfors, 2010; Navarro et al., 2012; Xu & Tenenbaum, 2007b). Each model correctly captures the empirical pattern in its own domain, yet when the central principles of both models are extended to the other domain – without making any special claims that something is different about generalization and categorization – they make the opposite prediction to each other. If this is true, it suggests that generalization *decreases* as experience of a category increases, but the probability of assigning items to that category *increases* with more observations. This is, to put it mildly, puzzling.

Why does this inconsistency occur? One somewhat dispiriting possibility is that the empirical evidence for one (or both) of the effects – however plentiful – is simply wrong. Alternatively, it is possible that both effects are real but are dependent on superficial properties of the task. One especially plausible possibility is that the source of the difference lies in the choice of dependent measure. Most categorization experiments use a forced-choice response while nearly all generalization experiments ask people to rate the probability of category membership. Though it is not obvious why this should drive different changes due to sample size, it is nevertheless a major difference between experiments that produce an expansion in categorization (as exemplified by, e.g., Nosofsky, 1988b) and those that produce tightening effects in generalization (as exemplified by, e.g., Navarro et al., 2012).

A more interesting possibility is the suggestion that there are *systematic* differences between the inductive problems posed by categorization and generalization tasks, and those explain the divergent patterns of results. Perhaps the mere fact that a generalization experiment presents people with positive examples of a target category whereas a categorization experiment displays examples of multiple categories is genuinely sufficient to produce a qualitative reversal in people's inductive inferences. Our discussion of the GCM and the Bayesian generalization model hints that this might be true, simply by virtue of the fact that Figures 1 and 3 both look like "sensible" models, but something seems to have gone awry with the model construction in both Figures 2 and 4. The "minimalist" adaptations that we have made to the GCM and the Bayesian generalization model both feel somewhat wrong for the problem they are ostensibly solving.

These considerations suggest that head-to-head comparison between a categorization task and a generalization task is required, taking care to keep everything else constant. In the experiments below, participants learned categories with different numbers of observations (either four or twelve, as in the simulations above). The measure of interest is whether their judgments expanded or tightened with the additional observations. Thus, Experiment 1 investigated inference when shown two categories, while Experiment 2 explored inference for one category. Within each experiment, we manipulated the task, asking separate participants either a forced-choice question about which category a novel item belongs to or a probability judgment question about how likely the novel item is to belong to the category. All other factors (e.g., appearance and nature of the stimuli, cover story, etc.) were kept constant.

Our results indicate that the sole important factor is the number of categories. In Experiment 1, where there are two categories, people's judgments *expand* regardless of the question they were asked. This behavior is predicted by the GCM but not the Bayesian generalization model. Conversely, in Experiment 2, with one category, people's judgments *tighten*, regardless of the question asked. This behavior is predicted by the Bayesian generalization model but not the GCM. In Experiment 3 we explore the possibility that these changes are caused by learners making different assumptions about the sampling process in the one- and two-category tasks. A direct instructional manipulation of sampling assumptions is sufficient to eliminate the category expansion effect in the two-category task. We conclude with a discussion of what these results mean about how categorization and generalization differ, and why the one-to-two-category shift should matter.

Experiment 1: Categorizing objects into two categories

Method

Participants. We recruited 500 participants on Amazon Mechanical Turk and collected data from 499 participants (the data from one participant was not saved). Of the 499 total participants, 23 were excluded because they had previously participated in similar online experiments run by our lab. An additional 94 were excluded for failing to meet a predefined accuracy threshold for non-critical test stimuli, described below. The remaining 382 participants were included in all analyses. Participants ranged in age from 18 to 79 (mean 32.6) with 43.5% being female. 72.0% of participants came from the USA, 23.6% from India, and all other countries less than 1%. People were paid $0.50 for their participation in the 8-minute experiment.

Design. Participants were randomly assigned to one of four conditions in a 2x2 between-subjects design. The first factor varied the number of exemplars from Category A that the participants were shown, either FOUR (N=209) or TWELVE (N=173). Category B consisted of four exemplars regardless of condition. The second factor manipulated the response elicitation method: people were either asked to provide FORCED CHOICE decisions in which they had to assign a novel item to Category A or Category B (N=228), or they were asked to rate the PROBABILITY that the novel item belonged to Category A (N=154).

Stimuli. Stimuli consisted of an outer rectangle that was 185 pixels wide and 110 pixels tall, with a vertical black line drawn on the interior of the rectangle. To assist people in making perceptual discriminations four evenly-spaced light grey vertical and horizontal grid lines were included within each rectangle. Example stimuli are shown in Figure 5.

Categories were defined by one dimension, but the nature of the dimension was randomly varied between participants: for some, the black lines varied by position, and for others, they varied by height. Stimuli were also left-right reversed for a random half of the participants. All analyses collapse across both of these factors, as none of them materially affected the conclusions.

Each stimulus varied along the one relevant stimulus dimension; we refer to this as the *value* of the stimulus. The value for line height varied from 5% to 95% of the height of the rectangle, and the value for line position varied from 5% to 95% of the width of the rectangle from the left edge. For clarity of exposition, we describe the rest of the experiment in terms of the condition in which the dimension varies by position and Category B values are encoded with higher numbers, even though both of these factors were completely randomized in the actual experiment.

Stimulus values were defined to match the values in the simulations in Figures 1 and 3. Category B, which was identical for all participants, contained four stimuli (as in the fourth row of Figure 5) with values of 73%, 77%, 85%, and 87%. For all participants, the range of values in Category A were identical, spanning from 21% to 43%. Participants in the TWELVE condition observed all of the odd-numbered stimulus values within that range in addition to these endpoints, resulting in twelve exemplars total. Those in the FOUR condition saw an additional two values randomly selected from between the endpoints, making four total.

Procedure. Participants were told that they were going to be shown a few example objects from two categories and then asked to make judgments about new objects while the examples remain on the screen. Participants making PROBABILITY response judgments were

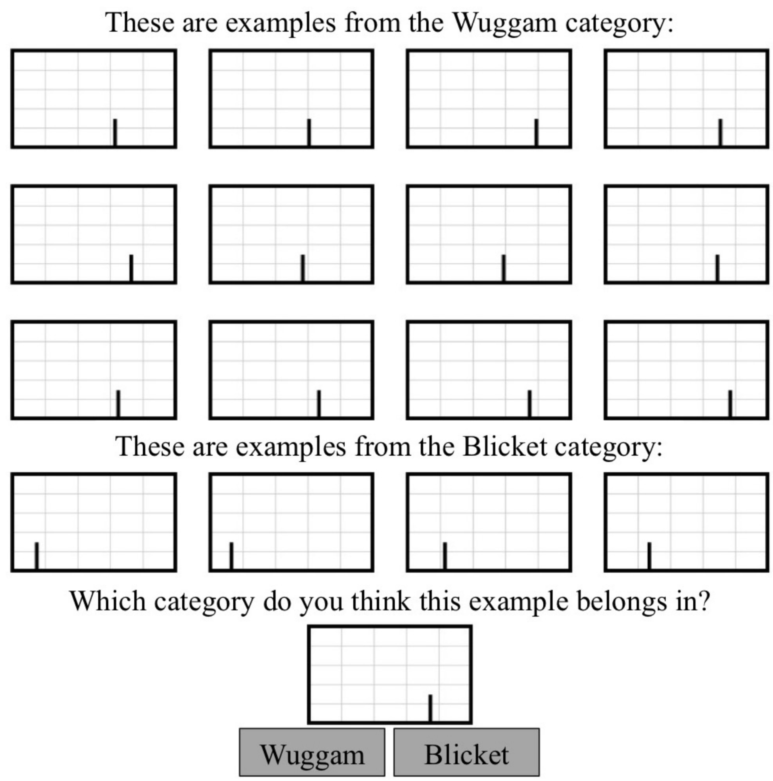

Figure 5. Sample stimulus display used in the two-category task (Experiment 1). The twelve training stimuli from Category A (labeled Wuggams) are in the top three rows, with the four stimuli from Category B in the fourth row. The single test exemplar from this trial is shown at the bottom of the figure along with two response buttons. The PROBABILITY condition looked identical except that instead of two buttons at the bottom people saw a slider with values ranging from 0% to 100% and a submit button.

instructed to "indicate how likely you think it is that this object belongs in the specified category." Participants making FORCED CHOICE responses were instructed to "indicate which category you think this object belongs in." Both conditions were required to answer a series of three check questions designed to make sure they understood the instructions and then proceed to a training and test session. Those that did not answer all of the questions right were returned to the instructions until they did.

Training. Training for Category A and B was simple and identical across all four conditions. In it, people saw exemplars from two categories defined according to a one-dimensional feature as described above. All of the stimuli appeared immediately, some as exemplars of Category A and some as exemplars of Category B, as in Figure 5. The stimuli in each category were arranged in a random order and each person saw a different assignment of two labels (Blicket and Wuggam) to the two categories. During the instructions people were told to "When you start the task you will be shown a few example objects from each category. Please take the time to study these examples and then press the Next button."

Testing. In order to minimize effects of memory, all training stimuli stayed on the screen for the duration of the test phase, which was very similar regardless of what question condition people were in. In all conditions, the set of test stimuli consisted of 19 exemplars that spanned the whole range of the stimulus dimension from 5% to 95% in steps of 5%.[4] The test stimuli were shown one at a time and in a random order. The next stimulus was shown directly after participants submitted their response by pressing a button.

What question people were asked during test varied by question condition. Those in the FORCED CHOICE condition were asked “Which category do you think this example belongs in?” and were then presented with two response buttons, one for each category. Those in the PROBABILITY condition were asked “How likely is it that this example is in the [Blicket/Wuggam] category?” and were then shown 21 radio buttons with labels going from 0% to 100% in steps of 5%. People in the PROBABILITY condition were always asked about the Category A label, as is typical in generalization experiments. In order to allow scope for capturing graded responding in the FORCED CHOICE condition, each test stimulus was presented once in the PROBABILITY condition and four times in the FORCED CHOICE condition. The conditions took similar amounts of time to complete.

Results

As mentioned above, 94 of the initial 499 people were excluded from the analysis for failure to achieve an accuracy threshold at test. This threshold was defined before analyzing any data and captured the intuition that if they understood the task and were trying, they should classify any stimulus with values between 21% and 43% as Category A (since that is the range of actually observed stimuli). One would expect that any stimuli with these values would be classified as A nearly 100% of the time, but in order to be as conservative as possible we set a threshold of 80%: those who classified the within-A items as not-A less than 80% of the time were excluded from the analysis (setting the threshold slightly higher or lower does not materially affect the results).

The left column of Figure 6 shows participant responses during the test phase for all stimulus values, with the top row showing performance in the FORCED CHOICE condition and the bottom row showing performance in the PROBABILITY condition. The overall pattern is consistent across conditions, with people more likely to correctly indicate that lower stimulus values are consistent with Category A and higher values are consistent with Category B. That said, the question of whether their responses are expanding or tightening with additional exemplars is only answerable upon examination of the critical stimulus values *between* the categories, shown in the right column of Figure 6. In both cases, especially in the FORCED CHOICE condition, there is a slight expansion away from Category A with additional training stimuli. Is this a statistically robust result?

To answer this question we compute a set of Bayes Factors that compare the relative posterior probability of different linear regression models that include different sets of factors. The data evaluated within these models consists of the critical stimulus values (45% to 70%, right column in Figure 6), with the two question conditions (PROBABILITY and FORCED CHOICE) analyzed separately.

[4]Due to a coding error, one of the extreme stimuli (either 5% or 95%) was not shown to 151 participants. Neither of these stimuli were within the critical region that we focus our analysis on.

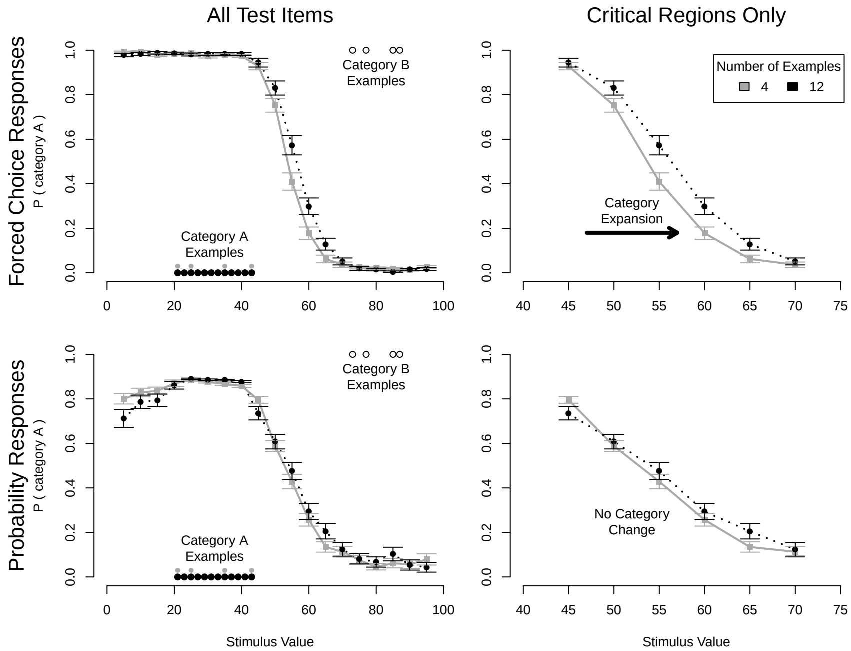

Figure 6. Human performance in the two-category experiment (Experiment 1). The top row shows the FORCED CHOICE condition: the proportion of answers selecting Category A in response to the question "Which category do you think this example belongs in?" for each possible stimulus value at test. The bottom row shows the PROBABILITY condition: the overall probability that the test item was rated as being in Category A. The graphs on the left show responses over the entire range of stimulus values; the right panels show responses for the critical range between the two categories. Points along the grey line indicate responses after seeing FOUR exemplars; the black points indicate responses after TWELVE. It is evident that participants who saw more observations in Category A *expanded* their responses away from Category A, especially in the FORCED CHOICE condition. This is consistent with the predictions of the GCM but not the Bayesian model of generalization.

We first consider performance in the FORCED CHOICE condition. Logically speaking, we are interested in evaluating the impact of three different factors on whether people define a stimulus with a given value as a member of Category A or not. As a sanity check, we should expect that **stimulus value** should have an impact on these judgments; people should be more likely to classify a stimulus located at 45% as Category A rather than one at 70%. The main variable of interest is whether the **number** of observations in Category A also plays a role in determining how people categorize the stimulus. We are also interested in whether there is an **interaction** between these two variables.

Table 1 shows the Bayes Factors (BFs) for each of those three models, with the model including only stimulus value as the baseline. Thus, the ratio reported for each model in the table reflects the BF for that model compared to the BF for the model containing only stimulus value as the predictor.[5] For the forced-choice judgments, it is evident from the

[5]We also ran a model containing only a random intercept for each individual. The next closest model,

Condition	Best Model	Model Performance		
		Value	Value + Number	Interaction
Category	Value + Number	1 : 1	**11.7 : 1**	0.08 : 1
Probability	Value	**1 : 1**	0.17 : 1	0.03 : 1

Table 1
Two categories: Comparison of how well three different regression models capture human performance in the FORCED CHOICE *and* PROBABILITY *conditions. All models are linear regression models with a random intercept for each individual. We consider three nested models: predictions based on stimulus value only, stimulus value and the number of observations, and one with both predictors as well as an interaction term. In the* FORCED CHOICE *condition, the preferred model is the one with both stimulus value and number of observations as predictors. This suggests that, in keeping with the GCM, people were more likely to classify an item as a member of Category A if there were more observations in Category A. In the* PROBABILITY *condition, the preferred model did not contain the number of observations. This suggests that probability judgments do not expand with category size in the same way (although the trend was still in that direction, rather than towards tightening).*

table that the most preferred model is the one that contains both stimulus value *and* the number of observations as predictors. This is evidence that people did in fact change their categorization probability when Category A had more observations in it. We can estimate how much this changed by examining the posterior estimates of the parameter values from the preferred model. They suggest that the eight extra training examples lead to a 7.5% increase in the probability of selecting Category A (95% CI: -2.5% to 12.7%) for test items between the two categories. This expansion in choice probability is consistent with the predictions of the GCM and not the Bayesian model.

The pattern of responses for participants in the PROBABILITY condition shows a similar trend, but is far less striking. This is evident quantitatively as well. As Table 1 shows, performing the same model comparison as before ends up favoring the model whose only predictor is stimulus value. In other words, the number of observations did not have a significant effect on people's answers to the probability judgment question. That said, the trend is in the same direction: it simply wasn't as strong, as is evident when examining the posterior estimates of the parameter values from the model that includes number of observations. It shows that the eight extra training examples lead to a 2.2% increase in probability judgment (95%CI: -2.2% to 6.6%).

with only stimulus value, was strongly preferred over the intercept only model ($BF > 10^{265} : 1$) so we abandon any further comparisons involving the random-intercept-only model. All of the models in Table 1 do also contain a random intercept for each individual. They were fit using the default parameters (Rouder, Morey, Speckman, & Province, 2012; Liang, Paulo, Molina, Clyde, & Berger, 2012) from the BayesFactor package (version 0.9.12-2) in R (version 3.3.0).

Discussion

The results from Experiment 1 are ambiguous in one sense, but very clear in another. It is not entirely clear whether an expansion effect was observed for both the FORCED CHOICE and PROBABILITY judgment tasks: taken at face value the results suggest an effect exists when people are asked to make force choice decisions, but disappears when asked to give probability judgments. While it might genuinely be the case that people categorize differently depending on the response measure, a more plausible explanation is that there is a weak category expansion effect regardless of the measure but forced choice judgements are more sensitive to changes than probability judgments. This ambiguity notwithstanding, one thing is very clear: there is no evidence for the category tightening effect predicted by the Bayesian model and shown in Figure 4. The results seem far more compatible with the GCM predictions shown in Figure 1.

Experiment 2: Generalizations about one target category

The superior performance of the GCM on a categorization task raises the possibility that it might also outperform the Bayesian approach on a generalization problem. Perhaps previous papers that found a tightening effect in generalization were false positives, or perhaps differences in experimental procedure can account for the difference in results. With this in mind, we conducted a second experiment in which people were shown examples from one category asked to make generalizations about new items, but in every other respect the procedure was the same as in Experiment 1.

Method

Participants. We recruited 500 participants on Amazon Mechanical Turk and collected data from 454 participants before the job posting expired. 45 participants were excluded because they participated in similar experiments. An additional 109 participants were excluded from all analyses because they failed the accuracy threshold at test (described in the previous experiment). The remaining 300 participants were included in all analyses. The participants ranged in age from 18 to 69 (mean 35.0) and 38.3% were female. 62.7% of participants came from the USA, 32.3% from India, and all other countries less than 1%. People were paid $0.50 for their participation in the eight-minute experiment.

Design. All details of the study, including the procedure and stimuli, were identical to the previous study. The only difference is that the stimuli were all from one category, which led to two minor differences in the text used to describe the situation. The training examples were now preceded by "These are examples from the category." The test question was slightly different as well: in the FORCED CHOICE condition, instead of asking "Which category do you think this example belongs in?" people were asked "Do you think this example is in the category?", while in the PROBABILITY condition they were asked "How likely is it that this example is in the category?" Participants were randomly assigned to the conditions with FOUR (N=125) or TWELVE (N=175) stimuli and were given either FORCED CHOICE response options (N=164) or a PROBABILITY judgment response (N=136).

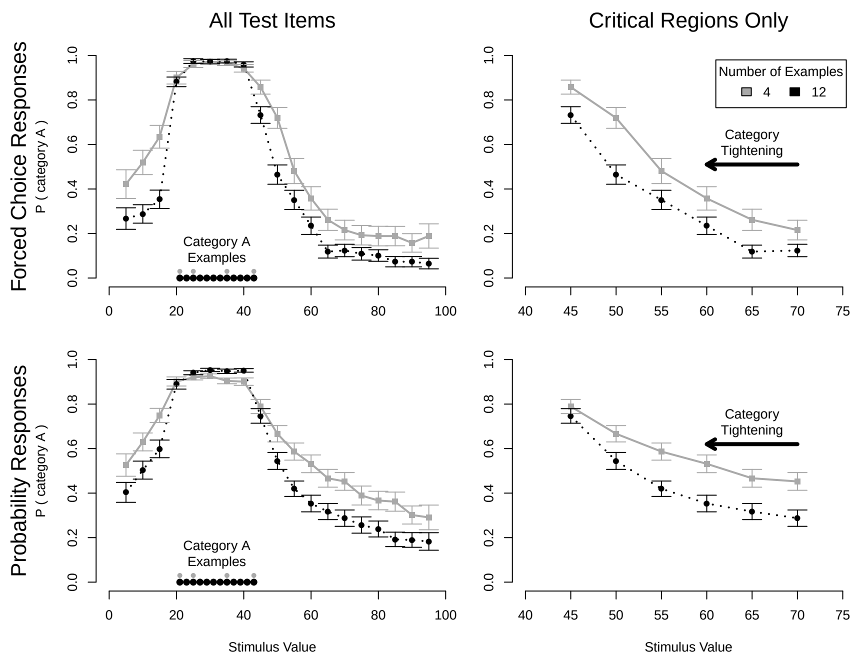

Figure 7. Human performance in the one-category experiment (Experiment 2). The top row shows the FORCED CHOICE condition: the proportion of answers selecting Category A in response to the question "Do you think this example is in the category?" for each possible stimulus value at test. The bottom row shows the PROBABILITY condition: the overall probability that the test item was rated as being in Category A. The graphs on the left show responses over the entire range of stimulus values; the ones on the right show responses in the critical range between the two categories from Experiment 1. Points along the grey line indicate responses after seeing FOUR exemplars; the black points indicate responses after TWELVE. It is evident that participants who saw more observations in Category A *tightened* their responses for Category A. This is consistent with the predictions of the Bayesian model of generalization but not the GCM.

Results

The left column of Figure 7 shows participant responses during the test phase for all stimulus values, with the top row showing performance in the FORCED CHOICE condition and the bottom row showing performance in the PROBABILITY condition. Unlike in the previous experiment, the overall pattern for both response types shows a *decrease* in probability of assigning test stimuli to category A when more training examples are added.

We quantify these effects by comparing the same set of linear models as before. As before, all models contain a random intercept for each participant and the posterior odds are estimated using the BayesFactor package in R. Furthermore, we only consider responses to the test stimuli from the same range as Experiment 1: 45% to 70%. Responses in the FORCED CHOICE and PROBABILITY conditions are analyzed separately.

Table 2 shows the Bayes Factors (BFs) for each of those three models; as before, the model including only stimulus value serve as the baseline. Thus, the ratio reported for

Condition	Best Model	Model Performance		
		Value	Value + Number	Interaction
Category	Value + Number	1 : 1	**7.7 : 1**	0.4 : 1
Probability	Value + Number	1 : 1	**50.1 : 1**	17.0 : 1

Table 2
One category: comparison of how well three different regression models capture human performance in the FORCED CHOICE *and* PROBABILITY *conditions. All models are linear regression models with a random intercept for each individual. We additionally consider the model that also contains stimulus value as a predictor, the one with stimulus value and number of observations, and one with both predictors as well as an interaction term. In both the* FORCED CHOICE *and* PROBABILITY *conditions, the preferred model contains both the stimulus value and number of observations as predictors. This suggests that, in keeping with the Bayesian model, people were more likely to classify an item as a member of the category if it had more observations.*

each model in the table reflects the BF for that model compared to the BF for the model containing only stimulus value as the predictor. It is evident from the table that the most preferred model in both the PROBABILITY and FORCED CHOICE conditions is the one that contains both stimulus value *and* the number of observations as predictors. This is evidence that people did in fact change their categorization probability when the category had more observations in it.

Examining the posterior estimates of the parameter values from the preferred model suggests that the eight extra training examples lead to a 13.7% decrease in the probability of including the stimulus in the category (95% CI: 2.1% to 23.8%) in the FORCED CHOICE condition and a decrease of 12.6% (95% CI is 4.0% to 29.2%) in the PROBABILITY condition. By contrast with the previous experiment, this tightening is consistent with the predictions of the Bayesian model rather than the GCM.

Discussion

The results in Experiment 2 are unambiguous. Regardless of whether people were asked to make forced choice decisions or to give probability judgments, increasing the sample size produced a tightening of the generalization gradients. As in Experiment 1, the estimated effect of the forced choice judgments appear to be more sensitive to the number of examples than the probability judgments. These results are consistent with Tenenbaum and Griffiths's (2001) Bayesian analysis and inconsistent with the predictions from the GCM. Although the GCM provided the better account of the categorization problem in Experiment 1, the Bayesian model provided a much better account of the generalization problem in Experiment 2. This occurred even though the two tasks employed the same stimuli, the same instruction set, the same recruitment procedure, and the same response elicitation methods.

Experiments 1 and 2 examine the effect of increasing sample size on both generalization and categorization tasks. Though the two tasks are conceptually closely related, the results indicate a clear difference. Experiment 2 showed that additional exemplars caused people to tighten their inductive generalization when learning about a single category. In contrast, Experiment 1 showed that such tightening was not evident in people's categorization decisions, and indeed the reverse effect was observed in the Forced Choice condition. If generalization and categorization are indeed closely related, then what might reasonably account for the differences observed?

The Bayesian model of generalization offers a possible explanation for these seemingly contradictory effects of increased sample size, in the form of the sampling assumption made by the learner. When items from only a single category are seen, as in generalization experiments, a learner might be justified in assuming that only items that belong in this category are being shown. This is equivalent to the strong sampling assumption from Tenenbaum and Griffiths (2001), and predicts that the learner should produce generalization curves that tighten as sample size increases, as observed in Experiment 2 with items from only one category are present.

In a categorization task, where items from more than a single category are present, it is less clear how the learner should assume items are sampled. One possible assumption would be the most conservative assumption [some citation here] about how items are sampled: that items are sampled independently of their category label. This is the weak sampling assumption (Shepard, 1987) and the prediction of how generalization curves change depends on how category frequency information is integrated. If frequency information is not incorporated into the decision rule then generalization curves do not change at all as sample size increases. If the decision rule depends on frequency (as in Equation 9), then generalization curves are predicted to expand as the relative frequency of the more frequent category increases.

Thus, the contrasting results found in Experiments 1 and 2 may be explained by a shift in sampling assumptions based on the number of categories presented. One way to test this account is by manipulating the beliefs learners have about why the frequency of the two categories are not equal. If participants are dynamically adjusting their sampling assumptions based on the category structure, then it should be expected that they can adjust the degree to which category frequency influences generalization. Experiment 3 tests this hypothesis using an extension of the two category design from Experiment 1.

Experiment 3: Manipulating sampling assumptions

This experiment replicates the Four exemplar condition of Experiment 1 and contrasts this against two alternative versions of the original Twelve exemplar condition, each designed to induce a different sampling assumption. The Twelve Random condition, where people are encouraged to believe that exemplars are selected at random (independently from category membership), is designed to induce a weak sampling assumption. In contrast, people in the Twelve Helpful condition are encouraged to believe that additional category A exemplars are chosen from the category by a helpful teacher. This condition is designed to induce a strong sampling assumption.

Importantly, we also seek to provide people with a plausible explanation as to why they are seeing more examples of one category than another in the two twelve exemplar conditions. By purporting to select items at random, the Twelve Random condition is designed to promote the belief that the different sample sizes are reflective of the true category base rates. Whereas, in the Twelve Helpful condition people are led to believe

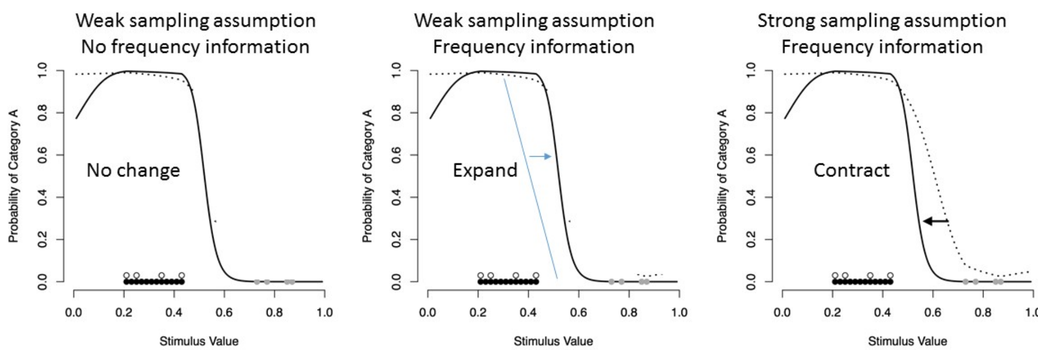

Figure 8. Three possible patterns of results in Experiment 3. The left panel is produced by assuming weak sampling and no frequency information in the decision rule. The middle panel is produced by a model that assumes weak sampling but includes frequency information when making a decision. The right panel is produced by a model that assumes strong sampling which inherently includes frequency information.

that the number of exemplars provided is simply an artefact of the experimental design (and therefore not reflective of the true base rates).

By allowing us to compare people's perfomance in the Four condition with each of the two alternative twelve exemplar conditions we are able examine the effect of additional exemplars on generalization, and to test whether the nature of the effect changes qualitatively depending on the learner's assumptions. In making such comparisons, there are three different patterns of results that we might reasonably expect, as illustrated in Figure 8. If a two category categorization task, by its nature, induces a weak sampling assumption (regardless of our cover story), we predict the following two effects. Firstly, people who observe twelve randomly sampled category A exemplars should attribute the difference in sampling frequency to a genuine difference in category base rates. As a consequence, these people should widen their generalizations away from category A when compared with people who see only four category A exemplars. Secondly, people who believe that the eight additional exemplars were sampled from the category should show no change in their generalizations toward category A relative to people in the Four condition. If, however, the Twelve Helpful sampling cover story is sufficient to lead people to believe that each item was strongly sampled from its respective category, then their generalizations should tighten toward category A relative to people who only saw four category A exemplars.

Method

Participants. We recruited 364 participants for this experiment via Amazon Mechanical Turk. Of these, 20 people were excluded from participation, having taken part in either of the previous experiments. No results were collected from 31 people who failed to complete the experiment. A further 15 people were excluded from further analysis for failing to reach the predefined accuracy threshold used in Experiment 1. Data from the remaining 298 participants were included in all subsequent analyses. Participants ranged in age from 18 - 68 (median age: 32), 39% were female, and 98% of participants were from the USA. Participants were paid $USD 1.25 for taking part in the 7 minute experiment.

Design. People were allocated at random to one of three conditions in an incomplete 2×3 between-subjects design. As with the two previous studies, the first factor varied the number of exemplars from category A that people were shown. The number of category B exemplars provided during training was fixed at four across all conditions. The second factor manipulated the cover story that people were given to explain how the training stimuli were selected. People in the Four condition ($N = 96$) were shown four exemplars from category B and four from category A, with no explanation offered for how these examples were chosen. Likewise, participants in the Twelve Helpful condition ($N = 99$) saw four exemplars from each category for which no explanation was offered. However, in addition they saw a further eight exemplars from category A which, they were told, had been selected from the category by a helpful teacher. In the Twelve Random condition ($N = 103$), people were told that 16 examples had been chosen for them at random. The "random" selection always consisted of the four category B exemplars, and twelve exemplars from category A.

Stimuli. The overall design of the study was based on Experiment 1. The stimuli were identical in appearance to those in the previous study (albeit that the images were allowed to scale to fit within the user's browser window in a manner that preserved the original aspect ratio). However, we restricted the presentation of the stimuli so that the black line within the stimuli varied only by position, with category A exemplars always represented by a line toward the left (the details of how the black line varied having made no material difference in the previous study). The same stimulus values were also used, as well as the same method of determining the subset of category A exemplars seen by participants in the Four condition.

Procedure. The experiment followed the same basic procedure as Experiment 1, the main difference being the initial explanation of the experiment given prior to the training phase. Participants in all conditions were told that the purpose of the experiment was to see how well they could judge between two categories of similar looking objects. Participants were then informed how examples would be selected. This explanation differed across the three conditions. People in the Four condition were told simply:

> We'll start by showing a few examples of each category, taken from our catalogues.

at which point the four category B and four category A exemplars were displayed on-screen. Participants in the Twelve Helpful condition were given this same introduction. However, after the initial exemplars were displayed, they were informed:

> **The computer has assigned you to experiment group «K8»**, so we're going to help you by showing you an additional «8» «Wuggams» **chosen by a helpful teacher** from our Wuggam catalogue.

People in the Twelve Random condition were told the following:

> **The computer has assigned you to experiment group «J16»**, so we'll start by **selecting «16» objects at random** from our catalogue. We'll classify the objects on-screen for you so that you have some examples to work with.

All subsequent instructions were identical across conditions.

As with Experiment 1 & 2, the training stimuli remained on-screen during the testing phase, and were annotated with a reminder of how the stimuli were chosen. Based on the assumption that forced choice decisions are more sensitive to changes than probability judgments, the response measure was limited to Forced Choice in all three conditions. Otherwise, the conduct of the test phase followed the procedure adopted in Experiment 1.

Results

The overall pattern of responses is consistent across all conditions, with people more likely to indicate that lower stimulus values were in category A, while higher values were in category B (Figure 9). Importantly, the pattern of responses is broadly consistent with that of Experiment 1 (Figure 6). Thus, we may use the results of the current study to meaningfully examine the question of whether the expansionary effect of additional exemplars observed in Experiment 1 was due to the assumptions on the part of the learner that the two category nature of the experiment induced.

Figure 9 shows the effect of additional observations on people's responses. The top row of the figure shows that people in the Twelve Random condition (responses shown in black) were more likely to classify test stimuli as belonging to category A, than people in the Four condition (response shown in grey). The bottom row tells a different story. It reveals that people in the Twelve Helpful condition (who were told that the eight additional exemplars had been selected from the category by a helpful teacher) responded little or no differently to people in the Four condition.

In order to further quantify these effects, we calculate posterior odds for three different linear models. In the Value Only model, predictions are based on the stimulus value only. In the Value + Number model, predictions are based on both stimulus value and the number of category A exemplars that participants observed. Lastly, the Interaction model extends the Value + Number model with a term that models an interaction between the two predictors. All models include the participant as a random effect. We analyse the two comparisons of interest separately. Firstly, we contrast responses from people in the Four condition against those from the Twelve Random condition. We do this by calculating Bayes factors to determine which of the three models best captures the data for the two conditions. Secondly, we contrast the Four and Twelve Helpful conditions, using the same technique. If the Value Only model best captures the data, then we can conclude that there is no effect of additional exemplars.

Bayes factors for the three models are shown in Table 3. For each model, the reported Bayes factor represents the ratio of the likelihood of the observed data under that

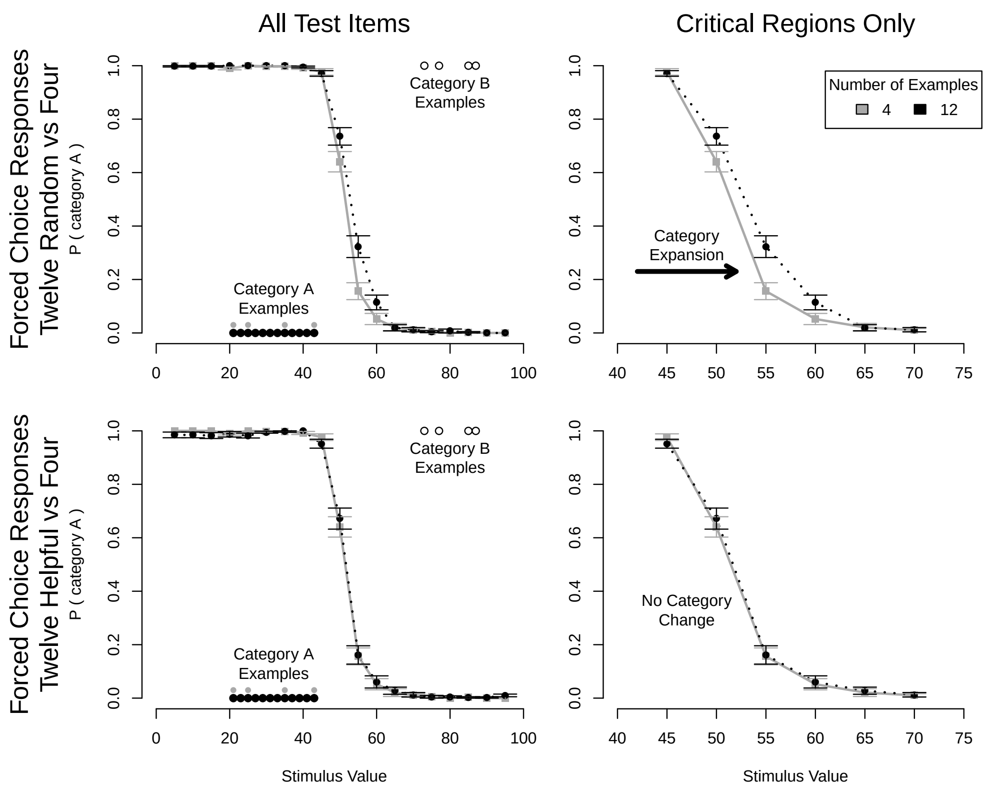

Figure 9. Human performance in the two category experiment with sampling assumption manipulation (Experiment 3). The graphs show the proportion of responses selecting category A in response to the question: "Which category do you think this example belongs in?" for each possible stimulus value at test. The graphs on the left show responses over the entire range of stimulus values; those on the right show responses for stimuli in the range between the two categories. Each row contrasts the performance of people who saw four category A exemplars (shown in grey) with one of the groups that saw twelve (shown in black): Twelve Random in the top row, and Twelve Helpful in the bottom row. It is evident that the effect of additional exemplars differed depending on condition. People who were told that all exemplars were selected at random from a collection of objects gave more responses in favour of category A within the critical region. In contrast, those who were told that an additional eight exemplars had been chosen by a helpful teacher, exhibited near identical responses to people who saw four exemplars only.

Contrast	Best Model	Bayes Factor (against VALUE ONLY)		
		VALUE ONLY	VALUE + NUMBER	INTERACTION
FOUR vs. TWELVE RANDOM	INTERACTION	1 : 1	2.5 : 1	**190 : 1**
FOUR vs. TWELVE HELPFUL	VALUE ONLY	**1 : 1**	0.10 : 1	2.8×10^{-4} : 1

Table 3

Comparison of how well three different linear regression models capture two different subsets of human response data in the two category experiment with sampling assumption manipulation. Each subset of the full data represents a different contrast of interest. Bayes factors are expressed as odds ratios against the VALUE ONLY *model, and are reported to two significant figures. Analysis of data from the* FOUR *and* TWELVE RANDOM *conditions shows that the* INTERACTION *model, which includes both stimulus value and number of observations as predictors, as well as an interaction term, best captures the data. In contrast, analysis of data from the* FOUR *and* TWELVE HELPFUL *conditions, shows that the* VALUE ONLY *model best captures the data. Taken together these results suggest that extra observations in category A are not sufficient to drive expansion of the category boundary. Additional assumptions, such as the representativeness of observed base rates, may also be required in order to drive expansion.*

model compared to the VALUE ONLY model. For the first contrast of interest (FOUR vs. TWELVE RANDOM), the INTERACTION model best captures the data. However, in the second comparison of interest (FOUR vs. TWELVE HELPFUL) the Bayes factors indicate that the VALUE ONLY model best captures the data.

Discussion

Experiment 3 shows a clear difference between the generalization curves produced by the random and helpful cover stories. Furthermore, these patterns align with the predictions of the extended Bayesian generalization model in which category frequency contributes to inference about sampling assumptions.

Let us first consider the TWELVE RANDOM condition, which was designed to persuade participants that items were sampled from a set of objects, and in such a way that relative differences in the number of items observed in each category were reflective of true differences in base rates. This condition is very similar to the non-cover story in Experiment 1 and, as before, the category A generalization curve expands as the frequency increases. This pattern is inconsistent with the Bayesian generalization model that assumes strong sampling, but is consistent with a Bayesian generalization model that assumes weak sampling and incorporates category frequency information into the decision. It is also consistent with the predictions of the GCM.

The TWELVE HELPFUL condition in contrast, was designed to persuade participants that the items were not selected randomly and therefore not in proportion to the base rate of the categories. Interestingly, this produced generalization curves that did not change relative to the FOUR condition. This pattern of results suggests that these instructions did not result in participants assuming that the items were sampled only from category A, which is strong sampling and thus would have predicted tightening of the generalization curve. Instead, the instructions resulted in participants assuming weak sampling and a decision rule that does not include information about category frequency, resulting in no change to the generalization curves.

The Twelve Helpful condition also demonstrates that increasing the relative frequency of a category in a categorization task may not be sufficient to drive category expansion. Instead, the effect of category frequency seems to be mediated by assumptions about how items are sampled. This poses an additional challenge for how to extend the GCM to account for these results.

General Discussion

Taken together, the results from Experiments 1 and 2 paint a very clear picture: the effect of increasing sample size is qualitatively different in categorization tasks and generalization tasks. In an inductive generalization problem, the learner is presented with positive examples that belong to a single category, and asked to determine whether novel items also belong to this category. In the generalization context, the Bayesian generalization model developed by Tenenbaum and Griffiths (2001) makes a clear prediction that is replicated here: the *size principle* implies that increasing the sample size without expanding the region of psychological space spanned by the exemplars should cause people's generalizations to tighten. This is precisely what Experiment 2 found.

In the categorization context, the opposite pattern holds. The exemplar model predicts that if we increase the sample size for one category without changing the region of psychological space that it covers, then the effect should be very small and in the opposite direction: a *base rate* effect should apply and as a consequence the categorization boundary should move *away* from the category. This is precisely what Experiment 1 found.

In one sense the results should be unsurprising: as a categorization model, the GCM outperforms the Bayesian model on categorization problems; and as a generalization model the Bayesian model outperforms the GCM on generalization tasks. However, unless one is prepared to give up on the notion that categorization and generalization are related inductive problems, this is not a terribly satisfying answer.

Experiment 3 attempted to test one possible explanation for the generalization differences between categorization and generalization: that learners in the two tasks make different assumptions about how the items are sampled. Without a method that can explicitly measure sampling assumptions, we turn instead to directly manipulating sampling assumptions via the cover story. The results are as clear as the previous experiments: participants not given a justification for the difference in category frequency show expansion of the more frequent category, participants given a justification beyond true category frequency show no expansion.

Addressing the inconsistency: the Bayesian generalization model

From the Bayesian perspective, a natural resolution to this inconsistency is to argue that participants bring different inductive biases to one-category generalization and multi-category categorization tasks. We might argue that in a generalization task, the *mere fact* that the experiment presents positive examples from a single category suggests that items are generated by sampling from items within the category. If this strong sampling assumption properly describes the learner's theory of the experimental task, then the size principle holds and we should expect generalization gradients to narrow as more data are received. When presented with a single category to learn about, this seems to be exactly what people assume is generating the items.

In a categorization task, it is not at all clear that the participant should assume a strong sampling model. Although not explicitly stated in these terms, the Bayesian categorization model outlined earlier in the paper is based on the assumption that the experiment generates stimuli via a two stage procedure. On any given trial the experimenter first selects a category (with probability given by the base rates); then having chosen a category, the experimenter chooses an exemplar *from* that category. This sampling procedure is the natural analog of the strong sampling model, as applied to a categorization task. It is not an inherently unreasonable assumption for a participant to make. Many real world categorization problems have a similar flavor, in which the categories are selected first, and then exemplars are sampled from those categories. The set of players competing in a soccer game come from such a sampling procedure: the two teams (categories) are first selected by the league and then players (item exemplars) from those teams define the set of players.

As reasonable as this sounds, it is by no means the only assumption that people might make about how a categorization task is designed. In a weak sampling model (as per Shepard's (1987) original description), the causal ordering is reversed: the world (or the experimenter) is assumed to have a fixed set of exemplars to choose from, and those exemplars are sampled independently of the category to which they belong. Again, there are many real world scenarios that seem to have this flavor. For example, when trying to guess the colors of vehicles, it seems natural to think that the world selects directly from the set of cars (item exemplars) without considering the color category to which they belong.

Critically, under a weak sampling assumption, the size principle does not hold and generalizations do not tighten (Tenenbaum & Griffiths, 2001; Tenenbaum, 1999). If people adopt this kind of sampling model in a categorization experiment – and assume that the experimenter chooses directly from the set of exemplars – then the Bayesian model does not predict a tightening effect. That is, a weak sampling categorization model would typically predict little to no effect of sample size or an expansion due to a base-rate effect in the prior probability of the category labels that changes with additional exemplars. What becomes less clear is how weak sampling alone (absent base rate information) can account for the difference between the random and helpful cover stories in Experiment 3.

Our point is not that the Bayesian model is incapable of capturing the effects seen in these experiments. Rather, our point is that it appears to account for one (tightening) quite naturally while the other (expansion) is less well theoretically motivated and more *post hoc*. Why should people prefer one sampling scheme over another in these different situations? Why, when shown a set of Daxes, do people appear to assume that the experimenter deliberately selected items from the *target category*, but when shown Daxes and Wugs they appear to assume that the items were selected from *the set of objects*? Since this difference arose even though the superficial aspects of the tasks were kept constant, the answer cannot have to do with the cover story, instructions, or response method used. Instead, a Bayesian model of how people perform both generalization and categorization must allow the learner to infer information not only about the hypotheses of the category structure, but also about how items are sampled.

We now turn to a consideration of the GCM. Can it be modified to predict both effects, and are those modifications more theoretically motivated and less post hoc than the modifications to the Bayesian model?

Addressing the inconsistency: the exemplar-based categorization model

The similarity-based exemplar framework offered by the GCM correctly predicts what happens in categorization tasks, but the straightforward adaptation of the model we implemented makes exactly the wrong prediction in the generalization problem. According to that model, as shown in Figure 2, the response strength for the category increases but the criterion against which it is compared does not, producing the expansion effect.

We can produce a different prediction by proposing an alternative decision criterion. Suppose that people automatically adjust the threshold to compensate for the rise in response strength for the target category. If, as in Nosofsky et al. (2011) where the threshold depends on list length in a recognition memory task, the threshold (ϕ) for accepting a new item increases as a linear function (m) of sample size (N) then Equation 4 would be replaced with

$$P(y \in c) = \frac{\eta(y,c)}{\eta(y,c) + mN\phi} \tag{12}$$

and the predictions of the model now become rather different. As Figure 10 illustrates, if the multiplicative factor for the relationship between threshold and sample size is one, the GCM now predicts a null effect of sample size in Experiment 2. By extension, of course, it is clear that the GCM *can* accommodate the qualitative tightening of generalization curves in the observed data perfectly well, simply by assuming that the scaling factor m is larger than one. Other, more fundamental, adaptations of the GCM framework to introduce dependencies between parameters and set size – such as changing the specificity parameter λ to account for changes in specificity due to the number of stimuli stored – can also be added to account for the tightening seen in the one-category generalization task. The process of constraining GCM parameter values to vary as a function of experimentally manipulated factors has been explored in a few papers including: constraining thresholds to vary as a function of list length in recognition memory (Nosofsky et al., 2011), and memory strength as a function of position in short-term memory tasks (Donkin & Nosofsky, 2012).

As with the Bayesian explanation, however, our point is not that the GCM is incapable of capturing both effects. Rather, it is that it appears to account for one (expansion) quite naturally while the other (tightening) is less well theoretically motivated and more post hoc. Why should the threshold or specificity parameter change as a function of sample size but only when no compelling reason is given for the frequency difference? Formalizing this relationship to make novel predictions about how thresholds (or other parameters) should change as a function of sampling assumptions and frequency is beyond the scope of this work but would require a full theory about how learners reason about sampling and frequency.

Sampling assumptions and category frequency

What then is the appropriate model of how learners incorporate information about the number of categories, the frequency of those categories, and their assumptions about

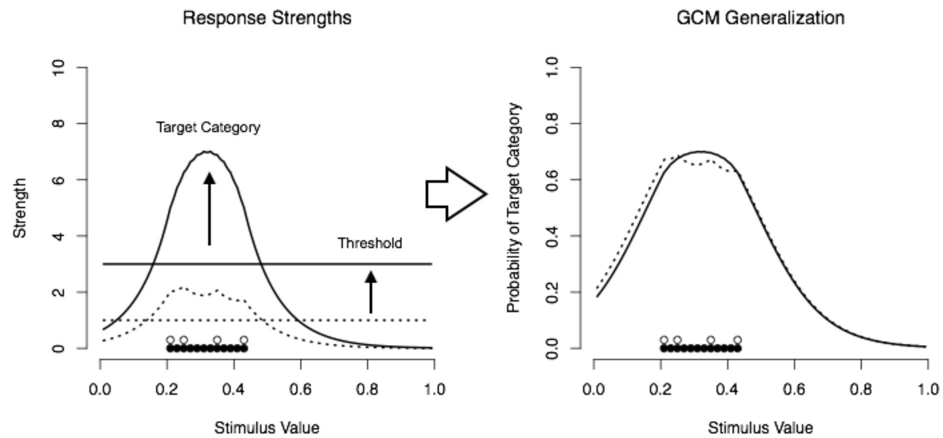

Figure 10. Modified generalization behavior in the GCM when the threshold scales linearly ($m = 1$) with sample size. The model now predicts that sample size should have no effect on people's willingness to endorse the category.

how categories and items are sampled? A complete answer to this question is beyond the scope of this article, however the results from the three experiments suggest the relationship involves two moving parts. First, learners appear to shift from category size dependent reasoning (strong sampling) in a one-category generalization context to a size independent reasoning (weak sampling) in a two-category categorization context. Second, the role of category frequency information in categorization is flexible and depends on the assumptions the learner makes about how items are sampled. However, treating categorization and generalization as fundamentally different tasks fails to capture the flexible nature of human cognition as it adapts and shifts beliefs as more information becomes available.

Acknowledgments

DJN received salary supported from ARC grant FT110100431 and AFP from ARC grant DE120102378 . Research costs and salary support for ATH were funded through ARC grants DP110104949 and DP150103280. KR was supported by an Australian Government Research Training Program Scholarship. Preliminary versions of this work were presented at the 48th and 50th Annual Meeting of the Society of Mathematical Psychology. We would like to thank Robert Nosofsky and two anonymous reviewers for their helpful comments on a previous version of this article.

References

Anderson, J. R. (1990). *The adaptive character of thought.* Psychology Press.

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

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*(1), 124 –150.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*(2), 154–179.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629–654.

Barsalou, L. W., Huttenlocher, J., & Lamberts, K. (1998). Basing categorization on individuals and events. *Cognitive Psychology*, *36*(3), 203–272.

Beyth-Marom, R., & Fischhoff, B. (1977). Direct measures of availability and judgments of category frequency. *Bulletin of the Psychonomic Society*, *9*(3), 236–238.

Boseovski, J. J., & Lee, K. (2006). Children's use of frequency information for trait categorization and behavioral prediction. *Developmental Psychology*, *42*(3), 500–513.

Breen, T. J., & Schvaneveldt, R. W. (1986). Classification of empirically derived prototypes as a function of category experience. *Memory & Cognition*, *14*(4), 313–320.

Donald, H., Joseph, C., Don, C., David, G., & Steven, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology*, *101*(1), 116.

Donkin, C., & Nosofsky, R. M. (2012). A power-law model of psychological memory strength in short-and long-term recognition. *Psychological Science*, *23*(6), 625–634.

Frank, M., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*, 360–371.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108—154.

Harris, H. D., Murphy, G. L., & Rehder, B. (2008). Prior knowledge and exemplar frequency. *Memory & Cognition*, *36*(7), 1335–1350.

Hawkins, G. E., Hayes, B. K., & Heit, E. (2016). A dynamic model of reasoning and memory. *Journal of Experimental Psychology: General*, *145*(2), 155–180.

Heit, E. (1996). The instantiation principle in natural categories. *Memory*, *4*(4), 413–452.

Homa, D., Burruel, L., & Field, D. (1987). The changing composition of abstracted categories under manipulations of decisional change, choice difficulty, and category size. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 401.

Homa, D., Dunbar, S., & Nohre, L. (1991). Instance frequency, categorization, and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 444–458.

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(3), 322.

Hsu, A., & Griffiths, T. L. (2016, 06). Sampling assumptions affect use of indirect negative evidence in language learning. *PLoS ONE*, *11*(6), 1-20.

Lewandowsky, S., Kalish, M., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1666–1684.

Lewandowsky, S., & Kirsner, K. (2000). Knowledge partitioning: Context-dependent use of expertise. *Memory & Cognition*, *28*(2), 295–305.

Lewis, M. L., & Frank, M. C. (in press). Understanding the effect of social context on learning: A replication of Xu and Tenenbaum (2007b). *Journal of Experimental Psychology: General*.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2012). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 410–423.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological Review*, *111*(2), 309–332.

Luce, R. (1963). Detection and recognition. In R. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Medin, D., & Schaffer, M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Myung, J. I., Pitt, M. A., & Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin & Review*, *14*(6), 1043–1050.

Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology*, *51*(2), 85–98.

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

Navarro, D. J., & Kemp, C. (under review). None of the above: A bayesian account of the detection of novel categories.

Navarro, D. J., & Perfors, A. (2010). Similarity, feature discovery, and the size principle. *Acta Pscyhologica*, *133*, 256–268.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114.

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

Nosofsky, R. M. (1988a). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708.

Nosofsky, R. M. (1988b). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 3–27.

Nosofsky, R. M. (2016). An exemplar-retrieval model of short-term memory search: Linking categorization and probe recognition. *Psychology of Learning and Motivation*, *65*, 47–84.

Nosofsky, R. M., Cox, G. E., Cao, R., & Shiffrin, R. M. (2014). An exemplar-familiarity model predicts short-term and long-term probe recognition across diverse forms of memory search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1524–1539.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological review*, *118*(2), 280–315.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 924–940.

Plummer, M., et al. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003). March* (pp. 20–22).

Polk, T. A., Behensky, C., Gonzalez, R., & Smith, E. E. (2002). Rating the similarity of simple perceptual stimuli: asymmetries induced by manipulating exposure frequency. *Cognition*, *82*(3), B75–B88.

Rips, L. J., & Collins, A. (1993). Categories and resemblance. *Journal of Experimental Psychology: General*, *122*(4), 468–486.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 726–731). Austin, TX: Cognitive Science Society.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Sanjana, N. E., & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. *Advances in neural information processing systems*, 59–66.

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

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 800–811.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. *Advances in neural information processing systems*, *11*, 59–65.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. *Advances in neural information processing systems*, *12*, 59–65.

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

Vandierendonck, A. (1988). Typically gradient in well-defined artificial categories. *Acta Psychologica*, *69*(1), 61–81.

Vong, W. K., Hendrickson, A. T., Perfors, A., & Navarro, D. J. (2013). The role of sampling assumptions in generalization with multiple categories. In *Proceedings of the 35th annual conference of the Cognitive Science Society* (pp. 3699–3704). Austin, TX: Cognitive Science Society.

Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative evidence? non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*, *81*, 1-25.

Williams, K. W., & Durso, F. T. (1986). Judging category frequency: Automaticity or availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 387–396.

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

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

Yang, L.-X., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 1045–1064.

Zaki, S. R., & Nosofsky, R. M. (2007). A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. *Memory & Cognition*, *35*(8), 2088–2096.