# Adding types, but not tokens, affects the breadth of property induction

**Belinda Xie (belinda.xie@unsw.edu.au)**
**Brett K. Hayes (b.hayes@unsw.edu.au)**
**Danielle J. Navarro (d.navarro@unsw.edu.au)**
School of Psychology, University of New South Wales
Sydney, 2052, Australia

**Abstract**

The extent to which we generalize a novel property from a sample of familiar instances to novel instances depends on the sample used. In these experiments, we are interested in two sample characteristics: number of types (discrete entities) and number of tokens (copies of the same entity) that share a novel property. Existing studies permit separate and conditional hypotheses about the effects of adding types and tokens, but no study has examined the effects of both variables on generalization stimuli varying in similarity. We find that adding types broadens generalization to similar stimuli, but tightens generalization to dissimilar stimuli. Adding tokens does not affect generalization, but adding repetitions that are framed as types produces some tightening. Implications for models of inductive reasoning are discussed.

**Keywords:** inductive reasoning, categories, concepts, Bayesian models

## Introduction

Imagine you are hiking with an ornithologist friend who points to several birds and tells you that these birds have *gabbro bones*. The next day, you hike alone and want to use your newly-gained knowledge to identify other birds with the same property. How will you decide which birds have gabbro bones? How far will you generalize? This is an example of a *property induction* problem, and the answers to these questions will depend on precisely which examples you were initially shown.

In this paper, we examine how people's inductive generalizations are shaped by two sample characteristics—the number of *types* and number of *tokens*. In this context, "types" are discrete entities that provide distinct information (e.g., a green parrot and a red parrot are different entities that represent two distinct types), whereas "tokens" are copies that provide redundant information (e.g., observing the same green parrot twice represents two tokens of the same type). In particular, our goal is to see whether types and tokens have analogous effects on the breadth of property induction.

### The effect of adding types?

Traditional models of inductive reasoning typically predict that adding types produces a *monotonicity effect*—increasing the number of premise exemplars that possess a property increases the likelihood of generalizing that property to a new conclusion exemplar within the same category (see Hayes & Heit, 2017, for a review). In the classic similarity-coverage model (Osherson, Smith, Wilkie, López, & et al, 1990), this effect arises because adding within-category exemplars increases the similarity between the premise category and a superordinate conclusion category that includes the premises.

In contrast, Bayesian models of property generalization (e.g., Navarro, Dry, & Lee, 2012; Tenenbaum & Griffiths, 2001) often predict that adding types also elicits a *non*-monotonicity effect: increasing the number of premise exemplars can *reduce* the likelihood of generalizing the property to novel exemplars, especially when those exemplars belong to a different category (e.g., Ransom, Perfors & Navarro 2016). When the reasoner observes more exemplars within a category that have the property, it strengthens the hypothesis that this category corresponds to the true extension of the novel property. The reasoner's beliefs thus converge to the smallest psychologically plausible category that contains the observed items. This phenomenon is known as the *size principle* (Tenenbaum & Griffiths 2001). Importantly, the size principle requires the reasoner to assume *strong sampling*, in which premises are selected from the set of objects that possess the property, rather than randomly selected.

Figure 1 illustrates this principle. For example, observing one green parrot with gabbro bones (the filled circle) results in moderately high generalization ratings for new exemplars that are highly similar to green parrots. Generalization ratings then decrease smoothly as a function of decreasing similarity (solid line). Conversely, observing four different green parrots (the empty squares) strengthens the belief that similar green parrots have gabbro bones, but *weakens* the belief that other dissimilar birds have gabbro bones (dashed line).

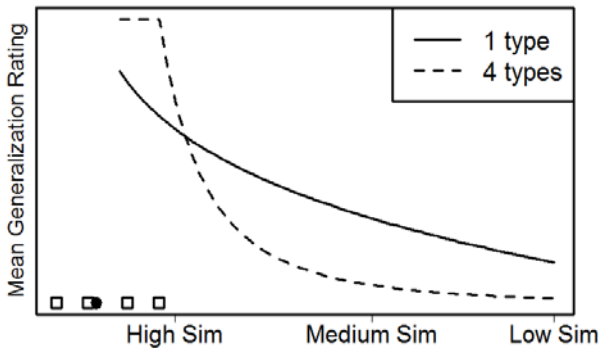


Figure 1: A Bayesian model of generalization predicts that increasing number of sample "types" from one to four will increase property generalization for high-similarity ("High Sim") test stimuli, but decrease generalization for medium- and low-similarity test stimuli.

The non-monotonicity prediction is significant because this phenomenon is not predicted by similarity-based models (e.g., Osherson et al., 1990). The empirical evidence for non-monotonicity however, has been mixed. When introducing the relevance theory of property induction, Medin, Coley, Storms, and Hayes, (2003) found evidence for non-monotonicity. For example, a property shared by Swedes, Finns, Danes, and Norwegians was less likely to generalize to Italians, than a property shared only by Swedes and Finns. This is because the additional types reinforced and made more relevant the property of "Nordic countries", subsequently weakening the strength of induction to Italians. By way of contrast, Fernbach (2006) failed to find evidence of non-monotonicity for biological categories such as lions and tigers (c.f. Ransom et al 2016). Participants generally preferred to generalize on the basis of a larger sample (more types) rather than a smaller sample.

An important limitation of these studies (Fernbach, 2006; Medin et al., 2003) is that they assessed generalization using a single conclusion category. This prevented them from testing the Bayesian predictions that adding types may increase generalization to instances that are similar to the sample, while also decreasing generalization to dissimilar instances. Hence, the first aim of our experiments was to examine the effect that adding types has upon property generalization to *a range of* novel exemplars that represent biological kinds and that vary in similarity to the sample of premise exemplars. To do so, we manipulated number of types of bird and flower exemplars (from one to four), and examined generalization to other birds or flowers representing four similarity levels (target, high, medium, and low). As illustrated in Figure 1, we predicted a monotonicity effect for high-similarity stimuli, but a non-monotonicity effect for low-similarity stimuli.

### The effect of adding tokens?

This first aim reflects the assumption of most property induction models that any additional exemplar added to an evidence sample is a discrete type that provides novel information. However, in both experimental and real-world contexts, new exemplars can seem very similar to old exemplars. In such contexts, it is not clear whether new instances will be perceived as new types or as new tokens of the same type. This distinction is important because, unlike types, new tokens provide no new information to the reasoner. For example, reading a second copy of a news story will not provide you with any additional information beyond the first copy (although it may increase your belief in its "truthfulness"; see Hasher et al., 1977). Thus, what informational value do people assign to new tokens, and how does this affect the breadth of generalization?

Little work has examined the effect of adding tokens on property generalization. A naïve interpretation of Bayesian theories of generalization is to treat all additional exemplars as discrete types, on the assumption that sampling the same entity twice provides new statistical evidence. However, other approaches are possible. Types and tokens are treated differently in Bayesian models of natural language production (Goldwater, Johnson, & Griffiths, 2006), object identification (Kemp, Jern, & Xu, 2009), and categorization (Navarro & Kemp, 2017). As such, it is not obvious a priori which approach is most appropriate to property induction.

To our knowledge, no study has examined the effect of adding tokens on property induction specifically, and the literature in related tasks has produced heterogeneous findings. In category learning, adding repetitions of the same exemplar can affect the likelihood that a new exemplar will be assigned to that category. For example, participants treated repeated presentations of the same color stimulus as separate instances when categorizing new colors (Nosofsky, 1988), and participants categorized new fish by relying on the exemplar fish that were presented most frequently (Barsalou, Huttenlocher, & Lamberts, 1998). However, in an artificial grammar learning task, Perfors, Ransom and Navarro (2014) found that generalizations about a novel grammar were sensitive to the number of distinct types, but—somewhat unexpectedly—were unaffected by the number of tokens.

With this in mind, our second aim was to examine how adding tokens affects people's willingness to make inductive inferences about a novel property. We manipulated number of observed tokens from one to four in Experiment 1, and two to six in Experiment 2. If participants treat repeated exemplars as types with novel informational value, we should see the same patterns as observed from our number-of-types manipulation. Conversely, if participants treat repeated exemplars as redundant tokens, then property generalization should be unaffected by additional tokens.

## Experiment 1

### Method

**Participants.** 1100 residents of the United States of America recruited from Mechanical Turk (MTurk). Data was lost for 107 participants due to a server overload, and 55 participants were excluded for failing an attention check question. The final sample size was 938 (48% female, median age = 34). Participants were paid $1.67USD for the ten-minute task.

Table 1: The 11 exemplar sets in Experiment 1 include every possible frequency table consisting of 4 or fewer tokens of the first type

| | *1 type* | *2 types* | *3 types* | *4 types* |
|---|---|---|---|---|
| *1 token* | 1 | 11 | 111 | 1111 |
| *2 tokens* | 2 | 21, 22 | 211 | |
| *3 tokens* | 3 | 31 | | |
| *4 tokens* | 4 | | | |

**Design and Materials.** The between-subjects design manipulated exemplar set, with each set varying in number of types and number of tokens. As shown in Table 1, there

were 11 exemplar sets, representing every possible way to distribute four or fewer tokens of the first type among four or fewer types. Exemplar sets are labeled in terms of the frequency table they correspond to. For example, in the "211" condition the sample was comprised of four instances: two tokens of the first type, one token of the second type, and one token of the third type (see Figure 2). Each participant was shown one exemplar set, chosen randomly.

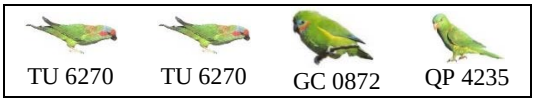


Figure 2: Exemplars shown in the 211 condition, in the bird trial of Experiment 1

**Procedure.** Participants completed one "bird trial" and one "flower trial" in counterbalanced order. In each trial, participants were asked to imagine they are researching how common a novel biological property (*gabbro bones* or *nelase enzymes)* is within a biological category (birds or flowers) on a newly-discovered island. Participants were told they would observe between one to four exemplars that have the property. The cover story made it clear that this sample could include types (e.g., photographs of different birds with the same novel property) or tokens (e.g., multiple photographs of the same bird). Type/token status was reinforced by the use of distinct or identical alphanumeric labels consisting of two letters and four digits (see Figure 2 for examples).

Participants were then presented with one of the sample exemplar sets from Table 1. Piloting showed that all sample stimuli were rated as having similar levels of typicality of their respective categories (bird or flower). Exemplars were displayed with their alphanumeric IDs, cumulatively, and from left to right on the screen. Each exemplar appeared onscreen for eight seconds before the next appeared.

After all sample exemplars were presented, participants were asked to generalize the novel property to other categories on the island (see Figure 3). There were seven generalization stimuli: one stimulus presented during training (hereafter the "target" stimulus); and two stimuli each of high, medium, and low similarity to the training stimuli, as determined in pilot similarity ratings. Each generalization stimulus was shown individually, in randomized order, with the instructions "Based on what you have learned so far, how likely is it that this bird has gabbro bones/flower has nelase enzymes?" Participants responded on a ten-point scale (where "1 = Definitely does not" and "10 = Definitely does").

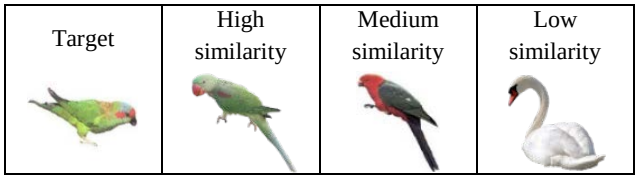


Figure 3: Example generalization stimuli for the bird trial

## Results and Discussion

Number of types and tokens were coded as categorical variables. Mean generalization ratings were coded as continuous variables, but the four generalization categories were discrete. We averaged generalization ratings across bird and flower trials (as they did not differ), and across the two stimuli within each of the high-, medium-, and low-similarity generalization categories. Analyses were performed using the BayesFactor package in R (Morey & Rouder, 2015) to compare ANOVA models.

Before discussing the between-subjects effects of adding types and adding tokens, we note that within-subjects generalization ratings decreased with decreasing similarity ($BF_{10}$ > 1000). The large positive Bayes factor indicates strong support for the alternative hypothesis of a difference between generalization categories, relative to the null hypothesis of no difference. This is an unsurprising finding predicted by both Bayesian and non-Bayesian models.

**Crossover effect from adding types.** We begin with the effect of adding types, in which all exemplar sets with the same number of types are grouped together (e.g., "1 type" includes the 1, 2, 3 and 4 conditions, "2 types" includes 11, 21, 22 and 31, etc.). As shown in Figure 4, adding types increased property generalization ratings for items with high similarity to trained items, but decreased generalization ratings for medium- and low-similarity items ($BF_{10}$ > 1000 in all cases). We describe this pattern of monotonicity for the high-similarity category and non-monotonicity at the medium- and low-similarity categories as a *crossover effect*.

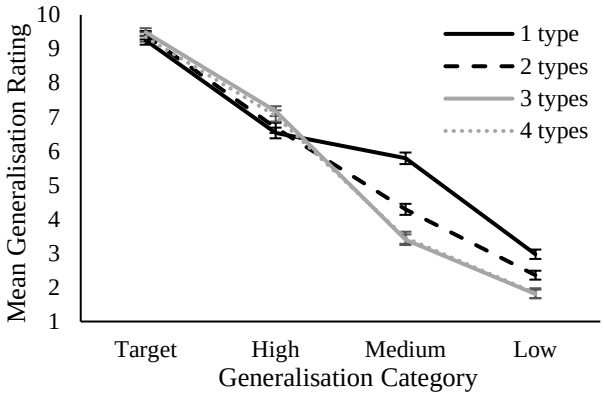


Figure 4: Adding types produces a crossover effect. In this and subsequent figures, error bars represent standard errors.

The effect of adding types was then examined for each level of numbers-of-tokens of the first type. The crossover effect was strongest when comparing 1→2→3→4 but was also observed for the 2→21→211 and 3→31 comparisons.

The monotonicity effect for the high-similarity category is consistent with our predictions about generalization to similar, within-category exemplars. Those who observed a sample with more types interpreted this as positive evidence

for the hypothesis that the property generalized to highly similar types—and thus gave higher generalization ratings.

Conversely, observing more types with the property was also seen as a signal that the property *only* applies to similar types—thus, decreasing generalization ratings for new types with medium and low similarity to the sample. This non-monotonicity effect is consistent with predictions made by Bayesian strong sampling models (Tenenbaum & Griffiths, 2001), and the size principle that these models instantiate.

**Null effect of adding tokens.** Figure 5 shows the effect of adding new tokens on property generalization, with all exemplar sets with the same number of tokens of the first type grouped together. Adding tokens did not affect generalization ratings to the target category ($BF_{10}$ = .008), nor to the medium- ($BF_{10}$ = .061), or low-similarity ($BF_{10}$ = .006) categories, but did cause tightening (i.e., lower generalization ratings) for the high-similarity category ($BF_{10}$ > 1000).

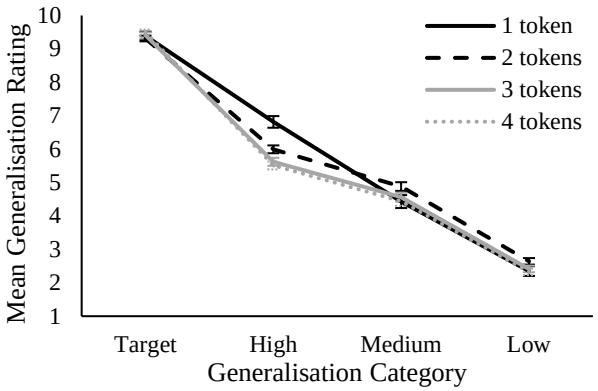


Figure 5: Adding tokens did not affect generalization ratings beyond the 1 token, 1 type condition.

Looking at the token effect for each level of number-of-types reveals that adding tokens ceased to have an effect beyond the 1→2 comparison. The added token from 1→2 tightened generalization for high-similarity ($BF_{10}$ = 7.1) and medium-similarity categories ($BF_{10}$ = 782). However, adding tokens produced a null effect with more than two tokens or more than one type ($BF_{10}$ < 1 in all cases). Thus, the main effect for the high-similarity category appears to be almost entirely driven by the 1→2 case. The overall null effect suggests that participants generally viewed repeated observations of an exemplar as having little effect on property generalization.

## Experiment 2

Experiment 1 found that adding tokens had little effect on generalization, suggesting participants were appropriately discounting the informational value of repeated exemplars. However, it is also possible that participants were simply ignoring the visually-identical repetitions, without considering whether repetitions should be treated as types or tokens. In order to clarify participants' ability to discriminate between types and tokens, Experiment 2 again examined the effect of adding tokens (from this point, we use the term "exemplars" to avoid confusion), but additionally manipulating whether repeated exemplars were explicitly described as tokens (i.e., repeated presentations of the same exemplar) or as types (i.e., new presentations of different exemplars). If participants are sensitive to the different informational value of types and tokens, generalization patterns will differ between the two groups. Specifically, adding observations should produce a null effect in the *repetition-as-token* conditions, and perhaps an attenuated crossover effect in the *repetition-as-type* conditions. Note that Experiment 1 used images of different parrots, whereas Experiment 2 used edited images of the same parrot, thus we do not expect a perfect replication of the crossover effect. We also predicted that the divergent effect on generalization will increase as the number of types or tokens increases—thus, we test an increased range of two to six exemplars.

### Method

**Participants.** 500 residents of the USA recruited from MTurk. Three participants had incomplete data, and 11 were excluded for failing the attention check question. The final sample size was 486 (52% female, median age = 33). Participants were paid $1.00USD for the six-minute task.

**Design.** The study used a 5 (number of observations: 2, 3, 4, 5, 6) x 2 (repetition type: repetition-as-token, repetition-as-type) between-subjects design, resulting in ten conditions. All participants were exposed to one type. The dependent variable was the same as in Experiment 1.

**Materials and Procedure.** This study used only bird stimuli (not flowers). Training stimuli were reflected and/or rotated versions of the first bird image used in Experiment 1 (i.e., the "Target" bird shown in Figure 3). Image transformations were used to increase the plausibility of the repetition-as-type cover story (i.e., that repeated exemplars represent discrete entities). Participants in repetition-as-token conditions were told that they may see multiple photographs of the same bird with the same ID number (as per Experiment 1). Participants in the repetition-as-type conditions were told that the same bird was never photographed more than once, and therefore repeated images represent different birds with different ID numbers. The training and test procedure was the same as in Experiment 1, except for an additional check of the repetition type manipulation. At the end of the experiment, participants in the repetition-as-token (repetition-as-type) condition were asked to rate "Based on the birds you saw, how much did you believe that some of the bird pictures were (not) repetitions of the same bird?" (1 = "Definitely not repetitions" and 10 = "Definitely repetitions").

### Results and Discussion

Responses to the repetition question confirmed that this manipulation worked as intended. Participants in the repetition-as-token conditions mostly believed birds were

repetitions of the same bird (mode = 10, $M = 7.66$, $SD = 3.13$), while participants in the repetition-as-type conditions mostly believed birds were not repetitions (mode = 1, $M = 4.88$, $SD = 3.31$). A Bayesian t-test provided very strong support for the alternative hypothesis that these means were different, $BF_{10} > 1000$.

Figure 6 shows the effect of adding exemplars separately for the two repetition type conditions. For repetition-as-token conditions (top panel), adding exemplars did not affect generalization at any similarity level ($BF_{10}$ ranged from .012 to .063), replicating the null effect in Experiment 1. For the repetition-as-type conditions (bottom panel), adding exemplars did not affect generalization to the target category nor did it have any effect for the medium-similarity and low-similarity categories ($BF_{10}$ ranged from .090 to .554). There was positive evidence for a difference in how people generalized to the high-similarity category ($BF_{10} = 5.50$). Visual inspection suggests this is a tightening effect with generalization ratings decreasing as the number of types increases from 2→3→4 exemplars, but no effect is observed beyond that. This tightening resembles the tightening observed at medium- and low-similarity categories in Experiment 1.

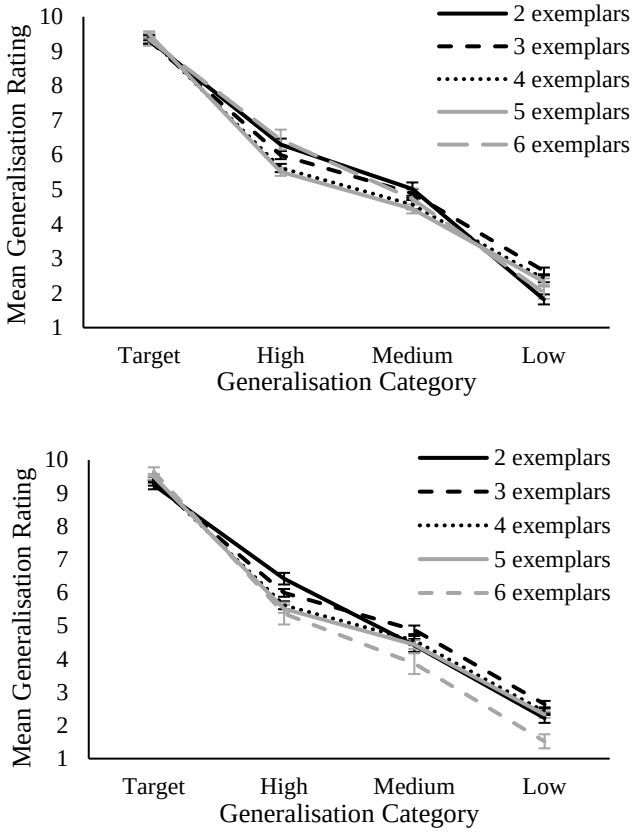


Figure 6: The effect of adding exemplars, in repetition-as-token (top panel) and repetition-as-type (bottom panel) conditions.

This pattern of results is different to Experiment 1, but may not be surprising given that the exemplars ostensibly depicting different birds in Experiment 2 were in fact visually identical. This lack of variability in the initial premise exemplar set may have caused the "high-similarity" generalization items in Experiment 1 to be treated as "medium-similarity" in Experiment 2. Regardless of the reason, Experiment 2 does provide suggestive evidence that the type/token status, even for identical-appearing exemplars, may affect how those exemplars are used in property induction. Adding exemplars when they are believed to be tokens of the *same* type had little effect, but adding the same exemplars as *different* types reduced generalization ratings—at least for some stimuli.

## General Discussion

Across two experiments we find that the breadth of property generalization can change quite substantially with the number of types, but not with the number of tokens. Adding types produced a crossover effect that is consistent with a Bayesian model of inductive generalization, while adding tokens produced a null effect that implies participants treated repeated exemplars as having little to no evidentiary value.

### The effect of adding types

In Experiment 1 we found that adding visually distinct types increased generalization ratings to high-similarity stimuli (i.e., a monotonicity effect) but decreased ratings to medium- and low-similarity stimuli (non-monotonicity). The former effect is consistent with the similarity-coverage model but the latter is not. To produce non-monotonicity effects in this model, one would require decreased similarity or decreased coverage, neither of which arises naturally without adding auxiliary assumptions to the model. In our studies, we chose exemplars to ensure that adding types did not decrease maximal nor average similarity between observed items and generalization targets. In contrast, the non-monotonicity effect emerges naturally within a Bayesian approach to inductive generalization that assumes strong sampling, and thus predicts the generalization patterns depicted in Figure 1 (Tenenbaum & Griffiths, 2001).

Although previous research has shown that sampling assumptions in inductive generalization are somewhat malleable (Ransom, Perfors, & Navarro, 2016), our results suggest that people rely on something akin to strong sampling in "typical" property induction scenarios dealing with types. This finding broadly mirrors the results from Medin et al., (2003), though does not agree with findings by Fernbach (2006). The inconsistency between the current results and Fernbach (2006)'s findings may arise from our use of multiple generalization stimuli along a similarity gradient, compared to Fernbach (2006)'s use of one generalization exemplar. In light of our findings that non-monotonicity only occurs when reasoning outside the category of exemplars, it is possible that his single conclusion exemplar (raccoons) was not sufficiently dissimilar from the various sets of premise exemplars to demonstrate non-monotonicity.

### The effect of adding tokens

The effect of adding tokens has not previously been investigated in property induction. In Experiment 1, adding tokens did not change generalization at any similarity level. This null effect is consistent with Perfors et al. (2014), but inconsistent with other studies in which token frequency affects categorization of novel instances (Barsalou et al., 1998; Nosofsky, 1988). Although these studies differ from ours in many ways, the most important differences are that: (1) we measured property induction as opposed to category learning, and (2) we clearly differentiated between types and tokens, leaving little room for participants to perceive tokens as types that provide novel information. On the first point, it is entirely possible that the relative effects of types and tokens differs in these different domains. However, it is also possible that participants in previous studies were simply unable to differentiate between types and tokens. Compared to our training set with a maximum of six exemplars representing four types, Barsalou et al. (1998) presented 30 exemplars of five different fish, while Nosofsky (1988) presented 48 instances of 12 colors. Perhaps the relatively lighter cognitive load placed on participants in the present experiments facilitated the "rational" ignoring of redundant repetition.

### The effect of type/token framing

To further investigate the extent to which participants attend to the relative informational value of types and tokens, in Experiment 2, we framed repeated exemplars as either new types, or new tokens of old types, while keeping visual information constant. Adding repetition-as-types decreased generalization to high similarity items, whereas adding repetition-as-tokens had no effect on generalization ratings (as per the null effect observed in Experiment 1). This suggests that people are sensitive to the difference in informational value between types and tokens even in this "pure framing" context—a sensitivity that many existing models do not explicitly accommodate. That being said, the effect size is modest and some degree of caution is warranted when interpreting this result.

### Future work

Similarity-based models of induction cannot account for the crossover effect of adding types. Bayesian models can, but they fail to predict the null effect of adding tokens. Our results therefore point to the need for a Bayesian model of generalization that assumes strong sampling for types, while also accommodating a different sampling process for tokens. This might resemble Goldwater et al., (2006)'s adaptor grammar model that allows different generative processes for generating new types compared to copying an old one, or Navarro, Dry, & Lee (2012)'s model that allows for some mixture of strong and weak sampling, depending on the evidence presented. Exactly how you generalize gabbro bones on your next hike is therefore determined by whether you were shown types or tokens—and cannot be adequately predicted by current models.

## Acknowledgments

BX is supported by a UNSW Scientia PhD Scholarship and an Australian Government Research Training Program Scholarship. The project was also supported by Australian Research Council Discovery Grant DP150101094 to BH.

## References

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

Fernbach, P. M. (2006). Sampling assumptions and the size principle in property induction. *Proceedings of the Cognitive Science Society* (pp. 1287-1292).

Goldwater, S., Johnson, M., & Griffiths, T. L. (2006). Interpolating between types and tokens by estimating power-law generators. *Proceedings of the 18th International Conference on Neural Information Processing Systems* (pp. 459–466).

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16(1)*, 107–112.

Hayes, B. K., & Heit, E. (2017). Inductive reasoning 2.0. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1459.

Kemp, C., Jern, A., & Xu, F. (2009). Object discovery and identification. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (pp. 925–933).

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

Morey, R. D., & Rouder, J. N. (2015). Package *BayesFactor*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

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. (2017). None of the above: A Bayesian account of the detection of novel categories. *Psychological Review, 124(5)*, 643–677.

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

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & et al. (1990). Category-based induction. *Psychological Review, 97(2)*, 185–200.

Perfors, A., Ransom, K., & Navarro, D. (2014). People ignore token frequency when deciding how widely to generalize. *Proceedings of the Cognitive Science Society* (pp. 2759-2764).

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

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