Adding types, but not tokens, affects property induction

Belinda Xie[a]*, Danielle J. Navarro[a], Brett K. Hayes[a]

[a]School of Psychology, University of New South Wales, Sydney, New South Wales 2052, Australia

Author Note

Keywords: Bayesian models; generalization; inductive reasoning; repetitions; samples; tightening

*Corresponding Author email: belinda.xie@unsw.edu.au

Acknowledgments: We thank Chris Donkin for helpful comments on the manuscript and guidance with modeling and statistical analysis. This work was supported by the Australian Research Council Discovery Grant [DP150101094] awarded to Brett K. Hayes. Belinda Xie is supported by an Australian Government Research Training Program Scholarship and the UNSW Scientia PhD Scholarship Scheme.

Declaration of interest: none.

Abstract

The extent to which we generalize a novel property from a sample of familiar instances to novel instances depends on the sample composition. Previous property induction experiments have only used samples consisting of novel types (unique entities). Because real-world evidence samples often contain redundant tokens (repetitions of the same entity), we studied the effects on property induction of adding types and tokens to an observed sample. In Experiments 1-3, we presented participants with a sample of birds or flowers known to have a novel property and probed whether this property generalized to novel items varying in similarity to the initial sample. Increasing the number of novel types (e.g., new birds with the target property) in a sample produced tightening, promoting property generalization to highly similar stimuli but decreasing generalization to less similar stimuli. On the other hand, increasing the number of tokens (e.g., repeated presentations of the same bird with the target property) had little effect on generalization. Experiment 4 showed that repeated tokens are encoded and can benefit recognition, but appear to be given little weight when inferring property generalization. We modified an existing Bayesian model of induction (Navarro, Dry & Lee, 2012) to account for both the information added by new types and the discounting of information conveyed by tokens.

Adding types, but not tokens, affects property induction

1. Introduction

When making inferences about novel cases or situations we typically rely on our knowledge of previously experienced samples. In the vast majority of work on inductive inference and judgment, these samples are composed of unique instances (see Hayes and Heit, 2017 for a review). For example, imagine you are hiking with an ornithologist friend who points to several parrots and tells you that they have the novel property *gabbro bones*. Your inferences about how far this property generalizes to other birds will be influenced by, among other things, the number of unique instances that you have observed in your sample (Osherson, Smith, Wilkie, López & Shafir, Eldar, 1990; Tenenbaum & Griffiths, 2001). Following Nosofsky (1988), we use the term "types" to refer to these unique instances.

But in everyday inference, the instances we encounter may not always be unique. Sometimes we will encounter the same instance several times. An advertisement for some product may be repeated throughout the day, an opinion piece may be shared by multiple social media users, or the same talking points may be repeated by numerous politicians. Back on the hiking trail, you may see the same parrot with gabbro bones multiple times as it flies in and out of the track. Nosofsky (1988) refers to such repeated presentations as "tokens", where, for example, four presentations of one exemplar constitutes four tokens. The type/token distinction is illustrated in the top two rows of Fig. 1, showing four parrot "types" (top row) and four "token" presentations of the one parrot type (second row).

How should token presentations influence inference? Logically, seeing an advertisement for a cleaning spray twice, compared to just once, adds no new evidence about whether the spray is effective. Similarly, repeatedly observing an instance that has a property should have little impact on inferences about how far the property generalizes. However, a number of lines of research suggest repeated information can significantly influence beliefs and behavior.

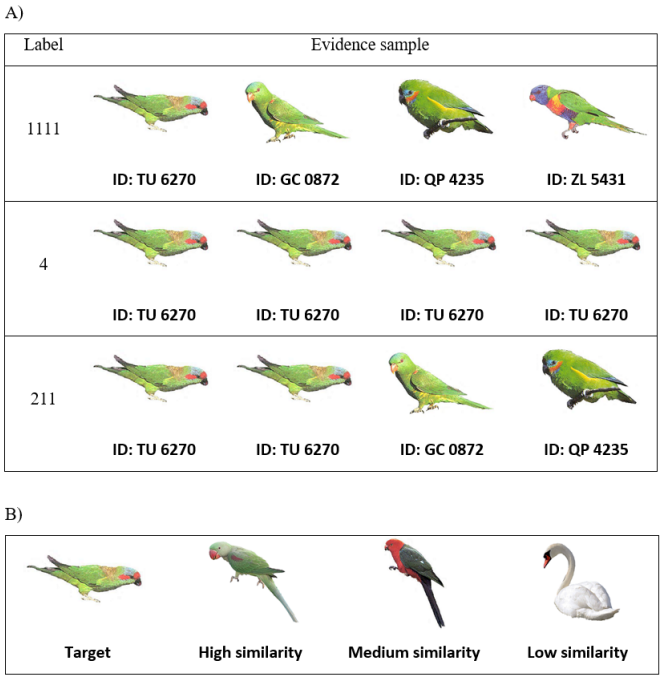

Figure 1. A) Three examples of evidence samples used in Experiment 1. Each evidence sample was labelled by a combination of numbers: the number of discrete digits represents the number of novel types shown in the evidence sample, and the numerical value of each digit represents the corresponding number of token presentations. For example, in *1111* (top row), four novel types (distinguishable birds with unique ID numbers) are each presented once. In *4* (middle row), a single token (same bird with same ID number) is presented four times. In *211* (bottom row), the first type is presented twice, while the second and third types are presented once each. B) Examples of generalization stimuli.

Over forty years ago, Hasher, Goldstein and Toppino (1977) described the "truth effect", in which repeating a plausible statement increased the belief in the truth of that statement. More recently, a large body of research (e.g., Barberá, Jost, Nagler, Tucker & Bonneau, 2015; Gvirsman, 2014; Sasahara et al., 2019) describes how repeating information in "echo chambers" can influence beliefs (Iyengar & Hahn, 2009), collective voter accuracy (Hahn, von Sydow & Merdes, 2019), and public policy (Jasny, Waggle & Fisher, 2015). Other experimental work demonstrates that people often 'double-count' information from dependent sources (Enke & Zimmermann, 2017; Gonzalez, 1994; Unkelbach & Rom, 2017), and can be swayed by "false consensus", or the repetition of a claim from a single source (Foster, Huthwaite, Yesberg, Garry & Loftus, 2012; Yousif, Aboody & Keil, 2019).

In the literature on inductive inference however, there is little work examining the effect of repeated token presentations. As discussed in detail below, models of induction typically assume that each observed instance constitutes a type – with each new instance conveying additional information about how far a property should be generalized. The current work therefore addresses the key question of whether inductive inferences are also affected by repeated presentations of the same instance. Our main aim was to compare the impact on property induction of adding token presentations to an evidence sample, compared to adding types. To do so, we manipulated type and token frequency in a number of property induction experiments. We then outline a formal Bayesian model that can address the effects of both added types and added tokens.

1.1. The effect of adding types in property induction

Early studies of property induction showed that observing more types that share a property increased property generalization to members of a more general category (Feeney & Heit, 2007; Osherson et al., 1990). For example, observing more parrots with gabbro bones increases the belief that all parrots have gabbro bones. This has been referred to as an effect of *premise monotonicity* (Osherson et al., 1990) or sample size

(Gutheil & Gelman, 1997). Subsequent work has shown that the effect of adding types is often more complicated. In particular, observing that many members of a specific category (e.g., green parrots) share a property can increase generalization to other members of the same specific category, but *decrease* generalization to members of other categories (e.g., other types of birds) (Hayes, Banner, Forrester & Navarro, 2019; Medin, Coley, Storms & Hayes, 2003; Ransom, Perfors & Navarro, 2016)[1].

Such generalization patterns are well explained by Bayesian models of inductive reasoning. Bayesian models (e.g., Navarro et al., 2012; Tenenbaum & Griffiths, 2001) treat property induction as a problem of selecting the most likely hypothesis about how far a property generalizes, given a sample of evidence. The learner begins with a broad prior hypothesis space consisting of many possible property generalizations (e.g., "green parrots have gabbro bones", "all birds have gabbro bones"). The learner updates beliefs in these hypotheses as new evidence is observed, with certain hypotheses becoming stronger and others becoming weaker. As well as changes in generalization due to additional types, Bayesian models can account for a range of inductive phenomena such as the effects of premise diversity (Hayes, Navarro, Stephens, Ransom & Dilevski, 2019), the impact of samples containing both positive and negative evidence (Lee, Lovibond, Hayes & Navarro, 2019; Voorspoels, Navarro, Perfors, Ransom & Storms, 2015) and generalization based on causal rather than categorical relations (Kemp & Tenenbaum, 2009).

A key assumption in many Bayesian models of inductive inference is that the observations are sampled from the set of objects that contains the to-be-generalized property (e.g., Navarro et al., 2012; Tenenbaum & Griffiths, 2001). Returning to the

---

[1] Our example here involves a sample of instances that are very similar to one another, as was the case with our experimental training stimuli. However, the predictions from Bayesian models about how additional types would affect generalization also apply to cases where the training instances are less similar to one another. For example, finding that a property was shared by parrots, turkeys and robins, rather than just parrots, would increase property generalization to other birds (monotonicity) but decrease or "tighten" generalization to other categories at the same level of abstraction as birds (e.g., mammals, fish) (see Ransom et al., 2016 for a demonstration of tightening with these kinds of stimuli).

previous example, strong sampling occurs when your friend on the hike only points out those birds *with* gabbro bones. This *strong sampling* assumption can be contrasted with a *weak sampling* assumption, where the observations are believed to have been selected randomly, regardless of whether they possess the property or not. Strong sampling appears to be the default assumption of learners in many experiments (e.g., Hayes, Banner et al., 2019; Ransom et al., 2016) and pedagogical contexts (Bonawitz et al., 2011).

Under strong sampling, if a category contains k items, the probability of observing an item is $1/k$, and for n objects (sampled with replacement), $1/k^n$. According to this *size principle*, as more types within some category (e.g., parrots) and with a given property are observed, the learner will favor narrower hypotheses (e.g., only parrots have gabbro bones) over broader hypotheses (e.g., all birds have gabbro bones). Adding types therefore reduces the likelihood of generalizing to dissimilar instances (e.g., an owl, an emu), because dissimilar instances are only included in larger hypotheses.

Fig. 2 shows an idealized illustration of the size principle and the "tightening" effect that follows: namely, decreased generalization to less similar stimuli. The x-axis here represents a one-dimensional similarity space. When generalizing from one type, generalization decreases gradually (solid line) as we move further along the x-axis (that is, as the generalization stimuli become less similar to the initial type). Generalizing from four types also decreases along the x-axis – but relative to generalization from one type, mean generalization ratings are lower for medium- and low-similarity stimuli. That is, generalization *tightens* with additional types. Fig. 2 also shows that adding types *increases* generalization to highly-similar stimuli – but only within a relatively narrow space of test-training similarity. Hence, in the current studies, the main focus is on whether we observe "tightening" in the form of reduced generalization to less similar stimuli as more instances are added to the training sample.

The psychological reality of the size principle is supported by experiments that demonstrate tightening when adding types to evidence samples in word learning (Xu & Tenenbaum, 2007), similarity judgments (Navarro & Perfors, 2010), and property

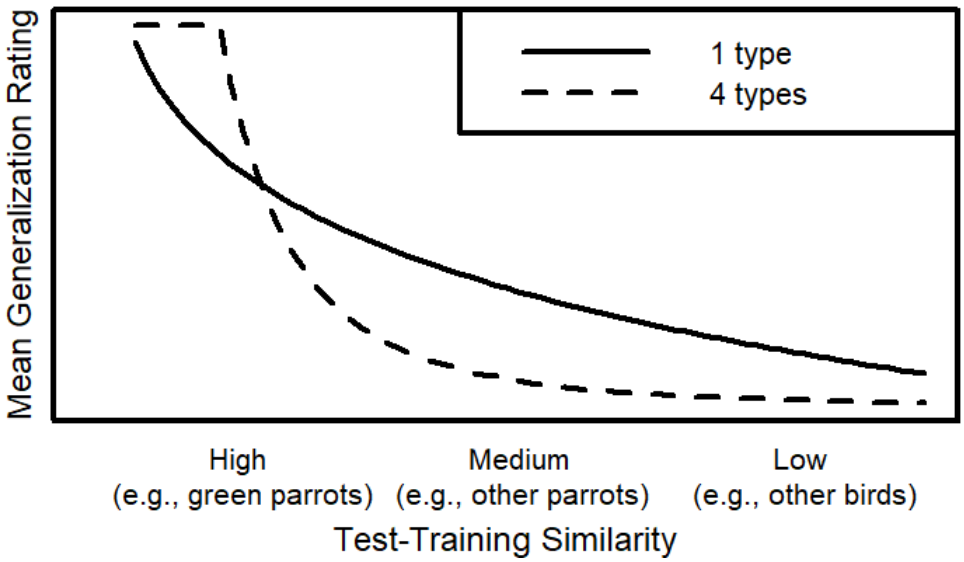

Figure 2. A Bayesian model of generalization predicts that increasing the number of types in a sample of evidence will tighten property generalization. This effect is illustrated in this idealized figure by the steeper generalization slope from four types (solid line) compared to from one type (dashed line).

induction (Navarro et al., 2012; Ransom et al., 2016; Sanjana & Tenenbaum, 2003; Tenenbaum & Griffiths, 2001). Tenenbaum and Griffiths (2001) for example, found that adding instances within a particular numerical range (presented as values that indicate "healthy hormone levels") reduced generalization to more distant values. Tightening has also been observed with more complex, multidimensional categories (Ransom et al., 2016; Sanjana & Tenenbaum, 2003). For example, Ransom et al. (2016) told participants about an animal that possessed a property (e.g., grizzly bears produce the TH-L2 hormone) and then asked participants whether the property generalized to a new animal (e.g., lions). Participants then learnt about a second animal that also possessed the property (e.g., black bears), and the generalization question was repeated. When participants believed that the sample instances were selected purposefully (i.e., strong sampling), they showed a decrease in generalization from the first to second trial, demonstrating a preference for smaller hypotheses. Conversely, participants who

believed the instances were selected randomly (i.e., weak sampling) increased property generalization from the first to second trial.

In the current work, we seek to provide a conceptual replication of tightening in property induction, while addressing two limitations of previous tasks. Previous studies have either only used unidimensional stimuli (Navarro et al., 2012; Tenenbaum & Griffiths, 2001), or only assessed generalization to a single conclusion category (e.g., samples of different sizes affect generalization to one category; Ransom et al., 2016). We seek to combine the advantages of both designs, by using naturalistic, multidimensional stimuli (cf. Ransom et al., 2016; Sanjana & Tenenbaum, 2003) and asking how the same sample can affect generalization to different categories representing different-sized hypotheses (Navarro et al., 2012; Tenenbaum & Griffiths, 2001).

1.2. The effect of adding token presentations in property induction

The more novel question addressed in the current work is how additional tokens affects property induction. Given the absence of previous work examining this question, we briefly consider relevant work in related cognitive domains. Prominent models of category learning such as Nosofsky's (1986, 1988) Generalized Context Model (GCM) assume that the frequency with which a token is presented will affect the way that novel stimuli are categorized. GCM assumes that subsequent categorization decisions are based on comparing a novel test item to learned items from candidate categories retrieved from memory. According to GCM, separate memory traces are stored for novel category types (e.g., two different parrots) and for token presentations (e.g., the same parrot encountered twice). Therefore, instances that have been encountered more frequently are more likely to be retrieved than less frequently encountered instances. Hence, categorization judgments will be biased in favor of the category membership of the high frequency instances.

To test this prediction, Nosofsky (1988) taught participants to classify color patches into two categories (roughly "pinkish" and "brownish" colors), after varying the presentation frequency of individual exemplars. In line with GCM predictions,

exemplars that were presented frequently during category learning were more likely to be classified correctly, and were judged as more typical of their respective color category, compared to exemplars that were presented less frequently (also see Barsalou, Huttenlocher & Lamberts, 1998).

This work shows that token presentation frequency can affect the way that exemplars are encoded and retrieved in categorization. However, it remains to be seen whether learners view token presentations as additional evidence in favor of an inductive hypothesis. A study by Perfors, Ransom and Navarro (2014) suggests that they do not. Participants were presented with ten separate types (unique sequences of letters or symbols) generated from the same underlying artificial grammar. These types were then either presented once, or repeated ten times. Perfors et al. (2014) found that token frequency had no effect on judgments about whether novel sequences belonged to the same grammar category as training instances.

These contradictory token frequency effects (cf. Nosofsky, 1988; Perfors et al., 2014) may seem puzzling. A possible resolution involves noting that the strongest effects of token frequency have generally been found when the task involved categorizing instances into one of two alternative categories (e.g., Nosofsky, 1988). By contrast, the Perfors et al. (2014) study involved a single category judgment of whether or not novel strings conformed to the grammar illustrated in a single training set. Arguably, property induction tasks have more in common with single-category learning tasks, than with the Nosofsky (1988) task, because they involve generalizing from a single evidence sample (cf. Hendrickson, Perfors, Navarro & Ransom, 2019). If this analogy is correct, then extrapolating from Perfors et al.'s (2014) findings, token repetition may have little impact on property induction.

We note that in some situations, additional presentations of old information do convey *some* information. To use our previous example, multiple people sharing an opinion article could serve to increase the credibility of the content. The fact that many individuals have independently chosen to share the content may be seen as adding to its evidential weight. Alternately, when there is a convergence of opinion between trusted

or expert sources, it is often normative to follow this shared advice (Harris, Hahn, Madsen & Hsu, 2016; Whalen, Griffiths & Buchsbaum, 2017).

In such situations however, it is hard to separate the learner's understanding of the relative value of repeated or novel information from social-cognitive factors such as beliefs about the intentions and reliability of the information source. Our goal was to examine the first of these issues in a learning context where additional token presentations added no objective information about property extension, and where the impact of social-cognitive factors was minimized.

A second way in which we simplified the learning environment was to make it easy for learners to know whether a given instance was a new type or a repeated token presentation, by adding individuating labels to each item. Outside of the laboratory, this distinction between "old" and "new" information is often not so clear. When several politicians repeat the same talking points, they may vary the order of presentation or other contextual details. When someone shares a previous post on social media, they may embellish it with their own comments. This complexity is also present in many previous experimental studies where repeated information is often mixed with new details. For example, in their study of false consensus, Yousif et al. (2019) presented multiple reports of the views of a single informant, but these repetitions were embedded in novel contexts (e.g., mock newspaper articles with different mast heads and including differing peripheral details). Hence it is unclear whether learners view each report as a repetition of previous reports or whether they were seen as adding new information. The goal of the current work was to create a more idealized learning environment so we could directly examine learner's understanding of the evidential contributions of old and new information, when the two forms of information are clearly discriminable.

1.3. The current studies

The current studies examined the effects on property induction of adding tokens to an evidence sample as compared with adding types. The studies extend previous work (e.g., Perfors et al., 2014; Tenenbaum & Griffiths, 2001) by manipulating both the

number of types and token presentations in the same experimental design. In each experiment, we presented participants with a sample of birds or flowers that possessed a novel property. This sample consisted of varying numbers of types (e.g., different birds with the property) and tokens (e.g., repeated presentations of the same bird with the property). Participants then made inferences about whether the property generalized to new birds or flowers that varied in similarity to the original sample.

In line with previous work (e.g., Sanjana & Tenenbaum, 2003; Tenenbaum & Griffiths, 2001), we hypothesize that adding types will produce tightening (i.e., decreased generalization to less similar stimuli). If people distinguish between new and old information in a normative way, then participants should not show this tightening effect with token presentations. On the other hand, if people fail to filter out tokens (as suggested by Enke & Zimmermann, 2017; Iyengar & Hahn, 2009; Nosofsky, 1988; Yousif et al., 2019), then such repetition should also lead to some degree of tightening. Note that these hypotheses naturally encourage the use of Bayesian statistical analyses that allow us to quantify evidence for and against the null hypothesis (e.g., a null effect of token repetition on generalization), compared to the alternative hypothesis.

2. Experiment 1

In this study, participants were presented with a small number of observations that possessed a novel property (training items), and asked to judge how likely it is that the novel property will generalize to each of seven new observations (test items) that varied in similarity to the training items (see Fig. 1B for examples). Across ten experimental conditions, we varied the number of types and token presentations, considering all possible type/token combinations in which the training items consist of no more than four stimuli.

2.1. Method

2.1.1. Participants. With ten distinct conditions and an initial target of approximately 100 participants per condition, we recruited a sample of 1000 adults via Amazon Mechanical Turk (MTurk). Participants were eligible if they resided in the

United States of America, and had a Human Intelligence Task (HIT) approval rate of at least 95%. However, data were not recorded for 115 participants due to exceeding a bandwidth limit on Google App Engine, where the experiment was hosted. Additionally, participants who failed at least one of the two attention check questions were excluded, leading to an additional 37 exclusions. The final sample comprised 848 participants ($n = 423$ female, $n = 417$ male, $n = 6$ other, median age = 34 years). Participants were randomly assigned to one of the ten conditions: see Fig. 3 for cell sample sizes. Participants were paid US$1.67 for the ten-minute task.

	Number of types			
Number of token presentations	1	2	3	4
1	Condition *1* (n = 80)	Condition *11* (n = 88)	Condition *111* (n = 94)	Condition *1111* (n = 73)
2	Condition *2* (n = 84)	Condition *21* (n = 89)	Condition *211* (n = 91)	
3	Condition *3* (n = 82)	Condition *31* (n = 79)		
4	Condition *4* (n = 87)			

Figure 3. Design of Experiment 1: There were ten experimental conditions, named according to the following scheme: the number of digits in each condition label represents the number of novel *types* shown in the evidence sample, while the numerical value of each digit represents the number of *token* presentations of that type. The total number of sample items shown in a condition (total token presentations of all novel types) is the sum of all digits. For example, in the *211* condition, four items were shown to participants; three instances were unique types and the first type was presented twice (two tokens). The number of participants for each condition are shown in parentheses.

2.1.2. Design and Materials. The ten conditions listed in Fig. 3 can be categorized by the number of distinct types (columns) and the number of token presentations of the first type (rows). Conditions falling along the diagonal correspond to cases where the total number of observations is held constant: for example, the *2*

condition and *11* condition both contain a total of two instances, but in the *11* condition these observations comprise two different types, whereas in the *2* condition, a single type is presented twice (i.e., there are two tokens). By necessity, a design such as this that includes all possible partitions of k or fewer tokens into k or fewer types will not be a fully factorial design with respect to the number of types and tokens, but such designs are not uncommon in the literature on categorization and inductive reasoning (e.g., Navarro & Kemp, 2017; Osherson et al., 1990; Ransom et al., 2016; Tenenbaum & Griffiths, 2001).

Screenshots from Experiment 1 and the complete set of stimuli are included in the Supplementary Materials (https://osf.io/mtbve/). Stimulus materials came in two forms, one using images of birds and the other using images of flowers. Examples of the bird stimuli used as training items and test items are shown in Fig. 1. The bird and flower images were gathered from a database maintained by the Sydney Thinking and Reasoning laboratory, and from copyright-free websites (pixabay.com, pexels.com). The image dimensions were approximately 180 x 180 pixels, and were selected after piloting showed no significant differences between these stimuli in their perceived typicality of the respective "bird" or "flower" categories.

All participants completed two versions of the task in randomized order, one using the bird stimuli and the other with flower stimuli. Each training image was presented with an identification (ID) code; a randomly generated combination of two letters and four digits. If two training items were supposed to represent different types, they were shown as discriminably different images of green parrots (or white flowers) with unique ID codes. If they were intended to represent repeated presentations of the same entity, the images were presented as copies of previously-seen ('old') images of green parrots (or white flowers) with identical ID codes (see Fig. 1 for examples).

The test items used to elicit generalization judgments were gathered from the same sources as the training sample. For each stimulus type (birds or flowers), the test set consisted of seven items; one familiar item taken from the training sample and two novel items at each of high, medium, and low similarity to the training sample. To

assess the level of perceived similarity between training and test items, we relied on similarity ratings made by an independent group of participants ($N = 158$) prior to the main experiment. Example generalization stimuli are shown in Fig. 1B; all other generalization stimuli are available in the Supplementary Materials (https://osf.io/mtbve/).

All analyses of test phase generalization ratings in this and subsequent experiments were based on averaged ratings for the four test item types (familiar, high-similarity, medium-similarity, low-similarity). That is, we averaged across the two items within each similarity level (e.g., averaged across the two high-similarity test items). We also averaged across the two stimulus types (birds and flowers), because preliminary analyses found few substantive differences in the generalization ratings for bird and flower items. The full set of non-aggregated data for all Experiments can be downloaded at https://osf.io/mtbve/.

2.1.3. Procedure. All Experiments were coded in jsPsych (v. 5.0.3 for Experiments 1-3 and v. 6.0 for Experiment 4; de Leeuw, 2015). Participants provided informed consent and basic demographic details before proceeding to the instruction phase. During the instruction phase, participants were asked to imagine that they were a scientist tasked with studying how common "gabbro bones" were among birds (or "nelase enzymes" among flowers) on an unexplored island. To make these judgments they would be shown a sample of items known to possess the novel property, which had been collected by research assistants. Participants were told that the research assistants worked independently of one another, and therefore it was possible that the same item could have been sampled multiple times (see Fig. 4 for a screenshot). Participants were told they could determine if two images corresponded to the same entity or different ones by checking the item ID code – different types had unique IDs, while token presentations of the same type had the same ID. To explain why the researchers had access only to such small samples, participants were told that the island was home to challenging terrain that made it difficult to capture larger samples. Four multiple-choice instruction check questions were included in the instruction set. Answering any

question incorrectly redirected participants back to reread the previous instruction page. Participants could not proceed until all instruction check questions were answered correctly.

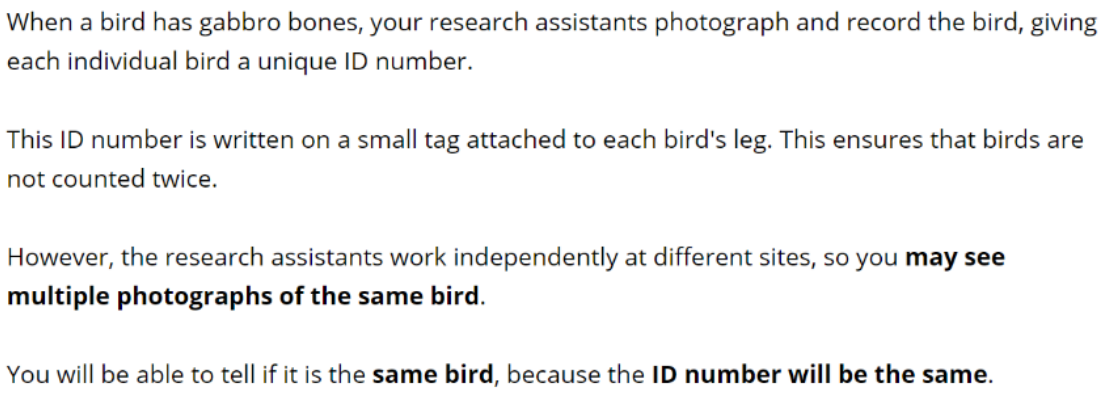

Figure 4. The onscreen instructions shown to participants to explain why repeated token presentations may occur, and how the ID numbers denote unique or repeated instances.

After the instruction phase was complete, participants were shown a training sample containing between one to four items that possessed the novel property (depending on the condition to which they had been assigned). Each item appeared on screen for eight seconds, then remained onscreen until all other items had been displayed. This cumulative presentation was used so that participants would not forget previous items, and made it easier for participants to make comparisons between training items. Once the last item had been onscreen for eight seconds, all items were removed and the generalization phase commenced.

As described previously, the test items consisted of seven generalization stimuli, which were displayed one at a time, in a randomized order for each participant. Each generalization stimulus was displayed above a generalization rating scale, and remained onscreen until the participant provided a response. For each test item, participants were asked to make their "best guess as to the likelihood that it also has gabbro bones (nelase enzymes)" on a scale from 1 to 10, where 1 was labelled *Definitely does not* and

10 was labelled *Definitely does*. After completing the generalization phase for the first set of stimuli (e.g., birds), the participant repeated the task for the other set.

2.2. Results

Analytic code (.R files) for all Experiments is available at https://osf.io/mtbve/. We calculated mean property generalization ratings separately for the four test-training similarity levels (familiar, high, medium, and low similarity) and the ten training sets (i.e., *11*, *2*, *21*, etc). The resulting generalization ratings as a function of added types and/or token presentations are shown in Fig. 5. Inferential analyses were conducted using the BayesFactor package in R with default priors (Morey & Rouder, 2015). These default Cauchy priors are appropriate for many designs and produce Bayes factors that are stable across changes in measurement scale (Rouder, Morey, Speckman & Province, 2012). To quantify the relative effects of number of types and token presentations, we compared a series of nested linear models and then conducted Bayesian analyses of variance (ANOVA) to explore any main effects. We obtained Bayes factors representing the relative odds of the data occurring under these models. The greater the Bayes factor (BF_{10}) is above 1.0, the greater the support for the larger model or alternative hypothesis, whereas the greater the inverse BF_{01} is above 1.0, the greater the support for the smaller, nested model or null hypothesis. As a guide, a BF_{10} from 1-3 provides weak evidence for the alternative hypothesis compared to the null hypothesis, from 3-20 provides positive evidence, 30-150 strong evidence, and greater than 150 provides very strong evidence for the alternative hypothesis (Raftery, 1995).

As one might expect, participants consistently gave ceiling-level generalization ratings for familiar test items (i.e., stimuli identical to those shown during training), regardless of the evidence sample (mean ratings ranged from 9.15 to 9.66). Moreover, the Bayes factor analysis suggested that these ratings were not affected by number of types ($BF_{01} = 50.82$) nor tokens ($BF_{01} = 122$). Given that the ratings for this condition are of little interest except as a check that participants recalled the information presented during training, the data from these test items are not analyzed further. The

results of the experiment for other test items are depicted visually in Fig. 5.

In each of Experiments 1 to 3 reported in this paper, there was very strong evidence for a main effect of test-training similarity (high vs. medium vs. low; all $BF_{10} > 1000$), compared to the null model with no effects. Generalization ratings decreased with decreasing similarity between test items and the training sample (see Fig. 5, 6, and 7). This result is unsurprising given the pervasive effect of similarity between sample and target items in previous studies of property induction (e.g., Heit & Rubinstein, 1994; Osherson et al., 1990). It also demonstrates that participants attended to the bird and flower pictures during training - not just the ID labels. With this in mind, our subsequent analyses of theoretically novel effects (i.e., adding types versus adding tokens) always include the main effect of test-training similarity before adding additional terms to the model.

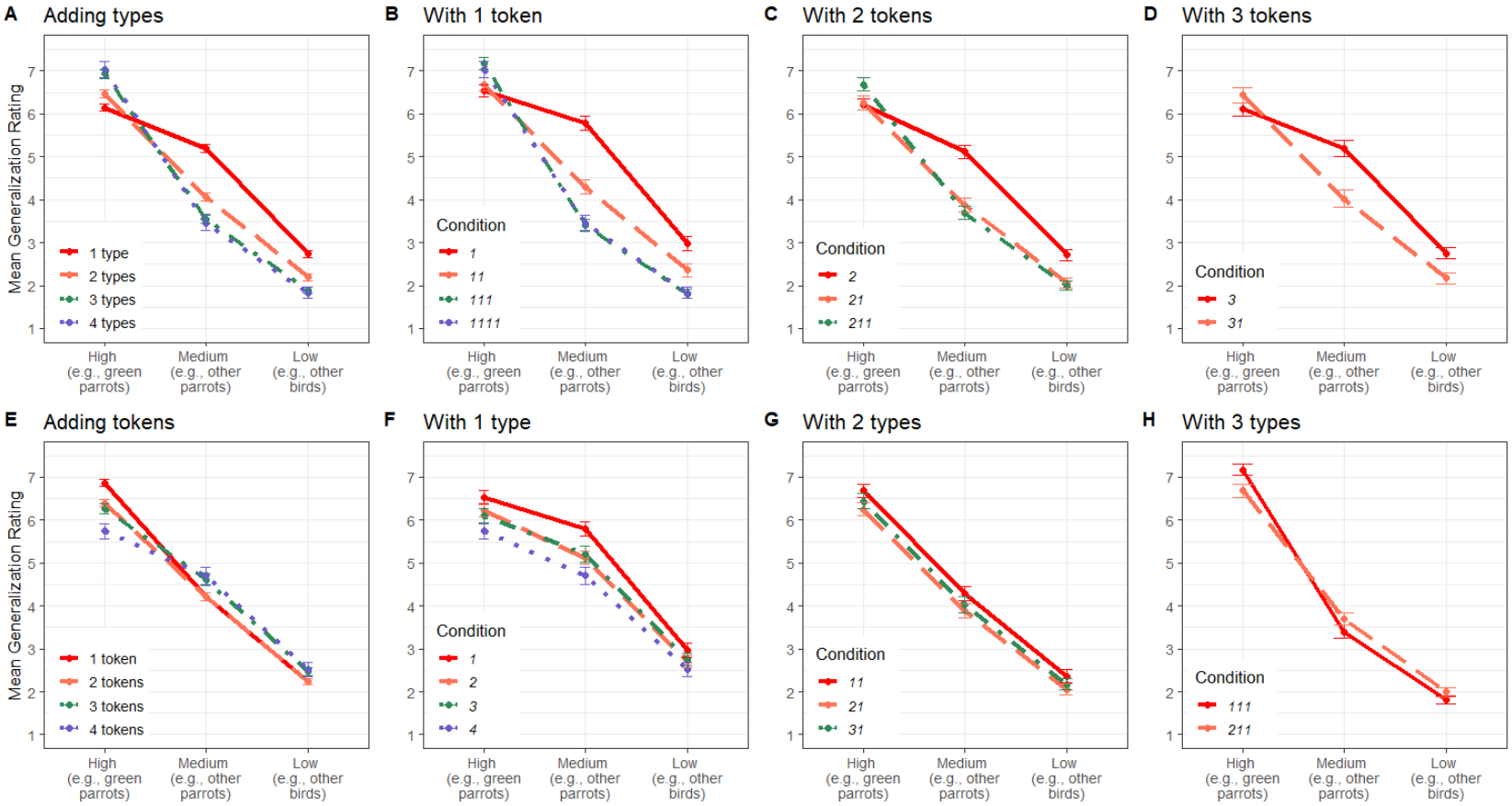

Figure 5. Experiment 1 Generalization Test Results: Panel A shows the effects of adding types averaged across variation in number of tokens. Increasing the number of types tightened generalization, by increasing generalization to high-similarity test items and decreasing generalization to medium- and low-similarity items. Panels B – D show that this tightening effect was consistent across evidence samples with one, two and three token presentations. Panel E shows the effect of adding token presentations averaged across variation in number of types. Increasing token frequency largely produced a null effect on property generalization. Adding tokens decreased generalization ratings when the evidence sample contained one type – see Panel F, but had no effect when there were two or three types – see Panels G and H. Error bars represent ±1 standard error around the mean.

2.2.1. The effect of adding types. We first consider the effect of adding new types (comparing conditions between columns in Fig. 3). The simplest way to do this is to aggregate data across the number of token presentations (i.e., collapse across all conditions within each column in Fig. 3) and compare the generalization ratings for conditions with one type, two types and so on. This is depicted in Fig. 5A: the horizontal axis plots test-training similarity and the vertical axis plots the mean generalization rating and its standard error. The four lines show the effect of adding more types. For the *high-similarity* test items people are most willing to generalize when shown four unique types, but are less willing to do so when the training set contains fewer types. However, for the *medium-* and *low-similarity* test items, this effect is reversed. This pattern follows the qualitative prediction from the Bayesian model shown in Fig. 2, and this interpretation is in agreement with the Bayes factor analysis: the data were much more likely under a model with two main effects of test-training similarity and number of types, and an interaction between these two variables, compared to the model with two main effects and no interaction ($BF_{10} > 1000$).

2.2.2. The effect of adding token presentations. We examined the effect of adding token presentations by comparing conditions between rows in Fig. 3. Adopting an analogous procedure to aggregate across number of types (i.e., collapse across all conditions within each row), we first compared all conditions with one token presentation, two token presentations, and so on. The overall effect of token frequency on generalization is shown in Fig. 5E. As before, the Bayes factor analysis provided strong evidence for the saturated model that contains main effects of test-training similarity, number of tokens, and the interaction term ($BF_{10} > 1000$) when compared to the next most plausible model.

Visual inspection of the results in Fig. 5E suggests that in this case, any effect is confined to the high-similarity item, for which adding token presentations tended to decrease generalization ($BF_{10} > 1000$, $\eta^2 = .05$). There is evidence against such an effect at medium and low levels of similarity: the Bayes factor analysis provided positive evidence for no effect of token presentations for medium-similarity items ($BF_{01} = 7.02$,

$\eta^2 = .01$) and low-similarity items ($BF_{01} = 10.03$, $\eta^2 = .01$).

2.2.3. A closer examination. In the analyses presented so far we have looked at the overall effect of adding new types, and the effect of adding new tokens of the same type. However, some care is required when interpreting these results because in both cases we collapsed across one of the two dimensions shown in Fig. 3, and have not looked at the ten conditions individually in search of interaction effects.

We first consider what happens when we disaggregate the effect of adding types shown in Fig. 5A. In panel B we plot the generalization ratings for those conditions with no repeated presentations, but where the number of types increases across the *1*, *11*, *111* and *1111* conditions. Panel C shows the same for the *2*, *21*, and *211* conditions, and panel D compares the *3* condition to the *31* condition. In all panels the same *qualitative* pattern is observed: generalization to high-similarity items increases with the number of types, but generalization to other items decreases. For each of these three cases we found strong evidence of an interaction between test-training similarity and number of types ($BF_{10} > 493$).

A similar approach can be applied to the effect of token presentations: the overall effect of number of presentations in panel E can be disaggregated into three special cases. In panel F we plot the effect of adding presentations when only a single type exists (i.e., the *1*, *2*, *3* and *4* conditions), panel G does so for the two-type cases (compares *11*, *21* and *31*) and panel H does so in the three-type case (comparing *111* to *211*).

In this case, we see a somewhat different pattern. The Bayes factor analysis suggests that adding tokens *does* have an effect when there is only a single type (i.e., $BF_{10} > 1000$ in panel F), but there is no such evidence in the other two cases ($BF_{01} = 2.18$ in panel G, and $BF_{01} = 10.94$ in panel H). Thus, when there was only one distinct type in the observed sample, seeing multiple presentations of that type reduced property generalization. Again, comparing across the three panels, there was weak to no evidence for an interaction between test-training similarity and number of token presentations ($BF_{10} < 2.81$).

2.3. Discussion

The results of this first study regarding the effect of adding types to an evidence sample were generally in line with the predictions of Bayesian models (like Tenenbaum and Griffiths, 2001) that incorporate the size principle (cf. Fig. 2). That is, we observed the tightening effect in which adding types decreased generalization to less similar instances. We also observed that adding types to the evidence sample increased generalization to highly similar test instances – an effect similar to previous demonstrations of premise monotonicity in induction (Feeney, 2007; Osherson et al., 1990).

However, when we look at the effect of adding *token presentations*, the empirical evidence is rather more ambiguous: while there does appear to be *some* effect of repeated token presentation, in these data it is confined to the (quite unusual) case where a single unique instance is presented multiple times. In the case where there are multiple types represented among the observed tokens – which would seem more representative of real-world inductive reasoning – we found no evidence for any effect of token frequency.

3. Experiment 2

Like the previous study, Experiment 2 also examined the effects on induction of adding types and tokens to an observed sample of evidence. In this case however, we used a factorial design that allowed for a more direct test of possible interactions between types and tokens. As in Experiment 1 we varied the number of types represented in the training set (one, two or three), but each type was represented by either two or four tokens. Using the notation from Fig. 3, we included the *2*, *22*, *222*, *4*, *44* and *444* conditions. Our goal in this experiment was to see if the key results of Experiment 1 would replicate with this slightly modified paradigm. Based on the earlier results, we expected to find a tightening effect with additional types, but little effect on induction of repeatedly presenting tokens.

3.1. Method

3.1.1. Participants. Participants were 561 adults recruited from Amazon Mechanical Turk (MTurk). The eligibility criteria used in Experiment 1 were again applied here, with the additional stipulation that those who had previously participated in Experiment 1 could not participate in Experiment 2. Two participants gave incomplete data, and 14 were excluded for failing the attention check question, leaving a final sample size of 545 (n = 238 female, n = 304 male, n = 1 other, median age = 32). Participants were paid US$1.17 for the seven-minute task.

3.1.2. Design and Procedure. The study used a 3 (number of types: 1, 2, 3) x 2 (number of token presentations per type: 2, 4) between-subjects factorial design, with participants assigned at random to one of the six conditions. For simplicity, we only reused the bird stimuli from Experiment 1, omitting the flower stimuli. As before, the dependent measures were mean generalization ratings for three levels of test-training similarity (high, medium and low similarity). The only difference in stimulus presentation from Experiment 1 is that we allowed images that represented token presentations of the same type to be reflected or rotated so that no two images were identical (images provided in the Supplementary Materials at https://osf.io/mtbve/). The ID codes remained identical, indicating that these tokens were repeated presentations of the same entity. We incorporated this change to reflect the fact that in real life we rarely see the same entity in precisely the same spatial context. The generalization stimuli were unchanged, and in all other respects, the procedure was the same as Experiment 1.

3.2. Results and Discussion

The test phase generalization ratings for Experiment 2 are shown in Fig. 6. Visual inspection of both figure panels suggests that there is again an effect of adding types. This was supported by the Bayesian statistical analysis: relative to a model containing only the effect of test-training similarity on induction, a Bayes factor analysis strongly favored the inclusion of a main effect of types ($BF_{10} > 1000$). Consistent with

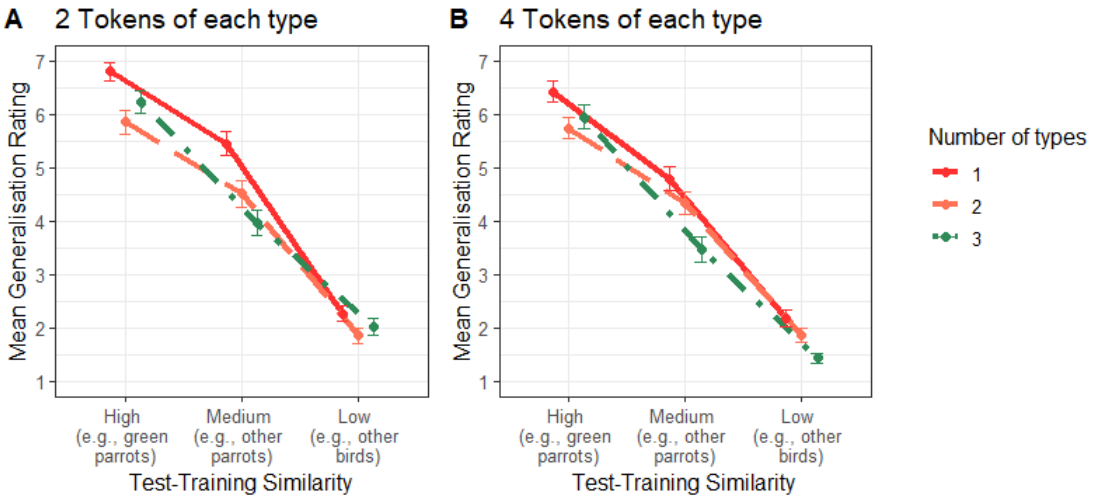

Figure 6. Experiment 2: Adding types decreased generalization to the medium- and low-similarity categories. Adding tokens had no effect on generalization, nor was there an interaction between number of types and token presentations. Note that points have been offset on the x-axis to improve readability and error bars represent ±1 standard error around the mean.

tightening, adding types decreased generalization for medium- ($BF_{10} > 1000$) and low-similarity items ($BF_{10} = 3.72$). Unlike Experiment 1, Fig. 6 shows that adding types did not lead to an additional "monotonicity" effect with increased property generalization to high-similarity stimuli. The absence of a monotonicity effect may be due to the fact that we included fewer unique types in this study (up to three) than in Experiment 1 (up to four). Fig. 5 shows that in Experiment 1, the strongest monotonicity effect for high-similarity test items was found when four unique types were observed during training. Nevertheless, the more standard tightening effect of decreasing generalization with less similar stimuli was replicated.

Consistent with Experiment 1, we found no statistical evidence for any effect of adding token presentations. The Bayes factor analysis found moderate evidence for the null model that contained main effects of test-training similarity and number of types, against an alternative model that also included an effect of number of token presentations ($BF_{01} = 9.36$). Notably, the factorial design in this study allowed for a

more direct test of an interactive effect of types and tokens. We found evidence against an interaction between number of types and number of tokens, with the Bayes factor analysis favoring the null model that did not include an interaction term ($BF_{01} = 9.00$). Recall that in Experiment 1, we found some evidence of decreased generalization when a single instance type was presented several times. For Experiment 2, we carried out an analogous analysis comparing generalization in the conditions in which a single instance type was presented two or four times (i.e., comparing *2* with *4*). This analysis failed to replicate the earlier trend with no clear evidence of an effect of token frequency on generalization ($BF_{10} = 1.55$).

The pattern of results from Experiment 2 provides further evidence that adding types to a training sample has a tightening effect on inductive generalization, while there is no corresponding effect of token presentations. Moreover, there was little evidence of an interaction between type and token presentations. The tightening effect of types is consistent with previous induction studies (Ransom et al., 2016; Sanjana & Tenenbaum, 2003; Tenenbaum & Griffiths, 2001), while the null effect of tokens agrees with work by Perfors et al. (2014). Given the novelty of this null token effect, we pursue the subtleties of token presentations further in Experiment 3.

4. Experiment 3

In Experiment 3, we consider situations in which the distinction between types and tokens is less obvious to participants. In Experiments 1 and 2, visual information alone was sufficient to determine whether a new token represented a new type or a repeated presentation of a previous one: types were represented by different images, whereas repeated presentations were represented by the same image, making the discrimination task relatively easy. However, in the real world, we often lack clear "objective" cues to tease apart these two situations. For example, two opinion pieces that support a certain policy may seem very similar, but this alone does not indicate that the two pieces are true repetitions – some additional information is needed. One might be more inclined to think that the opinion pieces represent distinct sources of evidence (and are hence more

akin to "new types") if they are published in two completely independent newspapers. If, however, the two newspapers are known to be owned by the same person, one might suspect a lack of independence and treat the two opinion pieces more like "repeated presentations" (cf. Gonzalez, 1994; Maines, 1996; Yousif et al., 2019).

In Experiment 3, we attempt to simulate this by presenting participants with multiple presentations of the same visual information – a single picture of a green parrot – but manipulating beliefs about whether the multiple presentations represent unique types or the same instance repeated a number of times. If participants remain sensitive to the higher informational value of novel types compared to repeated presentations – even when the stimuli appear identical – generalization patterns should differ between the two groups. If that is the case, we would expect that adding instances should produce tightening when the stimuli are believed to represent unique types ("type-label instances") but should produce a null effect when the stimuli are purported to be token presentations ("token-label instances").

4.1. Method

4.1.1. Participants. Participants were 500 adults recruited from MTurk. Eligibility criteria were the same as in the earlier experiments, and participants who completed either Experiments 1 or 2 were ineligible. Data was incomplete for one participant, and we excluded six participants who failed the attention check question. The final sample size was 493 (n = 253 female, n = 238 male, n = 1 other, median age = 33 years). Participants were paid US$1.00 for the six-minute task.

4.1.2. Design and Procedure. The study used a 5 x 2 between-subjects design, varying the number of instances presented (2, 3, 4, 5, or 6) and the manner in which those instances were labelled (as types or tokens). In all cases the visual stimulus was the same (i.e., one green parrot), though we allowed this image to be rotated or reflected as per Experiment 2. In the type-label conditions, the images were presented with different ID numbers, whereas in the token-label conditions the ID was identical for all images. This manipulation was also reflected in the cover stories (see Appendix

A for verbatim instructions): participants assigned to token-label conditions were told that their research assistants worked independently, and therefore the same item could have been sampled multiple times. By contrast, participants assigned to type-label conditions were told that their research assistants worked at the same site, and therefore the same item was never sampled more than once. After completing all generalization ratings, participants were asked – as a manipulation check – to rate how much they believed the birds they saw were repetitions of the same bird on a 10-point scale, where a response of 1 was labelled *"Definitely not repetitions of the same bird"* and 10 was labelled *"Definitely repetitions of the same bird"*. In all other respects the procedure was the same as Experiment 1.

4.2. Results and Discussion

4.2.1. Manipulation Check. Participants in the token-label conditions most frequently reported that bird pictures were definitely repetitions of the same bird (mode = 10, $M = 7.66$, $SD = 3.13$), whereas participants in the type-label conditions most frequently reported that bird pictures were definitely not repetitions of the same bird (mode = 1, $M = 4.88$, $SD = 3.31$). A Bayesian t-test supported the alternative hypothesis of a difference between means ($BF_{10} > 1000$). Thus, it appears that the label manipulation did elicit different beliefs about the nature of presented instances. Note that including data only for those participants who responded with 10 in token-label conditions or 1 in type-label conditions ($n = 184$) did not change the conclusions described in the following Section 4.2.2.

4.2.2. Generalization Ratings. The results of Experiment 3 are displayed visually in Fig. 7. On average, generalization ratings in the type-label conditions are lower ($M = 3.75$, $SD = 2.46$; Fig. 7A) than the corresponding judgments in the token-label conditions ($M = 4.25$, $SD = 2.46$; Fig. 7B). A Bayes factor analysis is consistent with this observation: the data are much more likely according to a model that includes effects of test-training similarity and label ($BF_{10} > 1000$) relative to a model containing only the effect of test-training similarity.

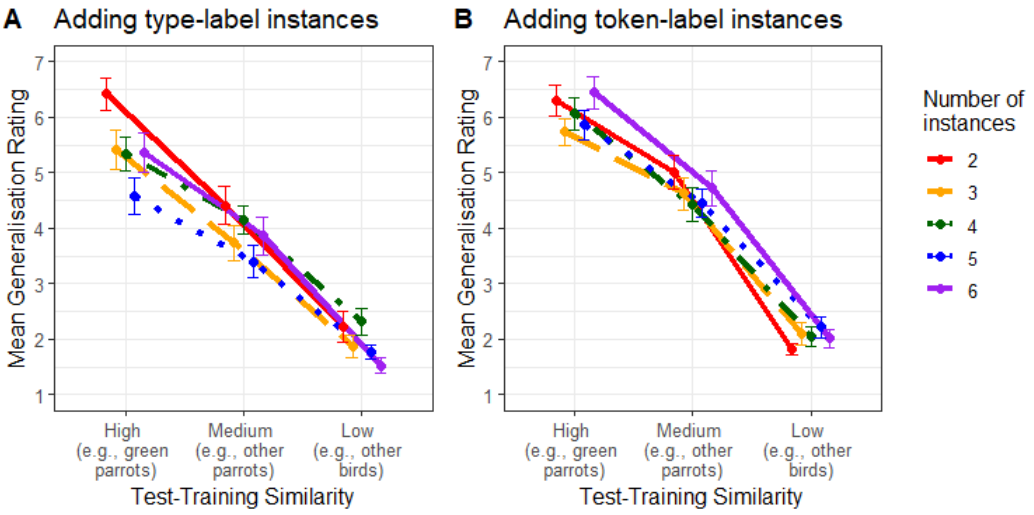

Figure 7. Experiment 3: A) Adding type-label instances decreased generalization to high-similarity stimuli. B) Adding token-label instances had no effect on generalization. Note that points have been offset on the x-axis to improve readability and error bars represent ±1 standard error around the mean.

What is the effect of increasing the number of items in the training set? According to the hypothesis that people treat types and tokens differently, we should expect an interaction effect: increasing the number of instances should decrease generalization in the type-label condition, and not affect generalization in the token-label condition. Visual inspection of Fig. 7 suggests that this is supported by the data, and a Bayes factor analysis is again in agreement with this intuition, though the evidence is only of moderate strength. The Bayes factor for the inclusion of the interaction between label and number of instances is $BF_{10} = 8.19$ relative to a model including only the main effects of label, number of instances and test-training similarity. Moreover, the interaction effect has the form we expected. As predicted, adding new instances that were labelled as token presentations (panel B) had no effect on generalization ratings ($BF_{01} = 12.13$).

Turning to the effect of increasing instances in the type-label condition (panel A) we again find results that broadly agree with predictions of tightening, but caution is warranted. Generally, increasing the number of instances labeled as types resulted in

lower overall generalization. Unlike the previous studies, we observed the largest effect of adding types for *high*-similarity stimuli ($BF_{10} = 5.50$), but little evidence for any effect of adding types for medium- ($BF_{01} = 7.81$) or low-similarity items ($BF_{01} = 1.80$). Moreover, the five-instance condition did not appear to follow the same ordinal pattern as the other conditions (i.e., it should have produced the lowest generalization ratings). This may simply be sampling error, and we are wary of running too many post-hoc analyses, so we merely note this unusual property in the data.

Experiment 3 supports the previous conclusions from Experiments 1-2 that adding new types can have a substantial effect on how people generalize, whereas repeated presentations of instances within an evidence sample have little effect on property generalization. Experiment 3 is additionally noteworthy because it showed that the differential effects of types compared to tokens apply even when the stimuli are visually identical, but *believed to be* either unique instances or repeated instances.

5. Experiment 4

Experiments 1 to 3 established that repeating instances in a training sample (i.e., adding tokens) has little systematic effect on inferences about property generalization. In Experiment 4, we distinguish between two possible mechanisms by which learners "filter out" such repetitions in the inference process. The first is that learners simply ignore repeated information; the psychological equivalent of 'covering one's eyes' to subsequent presentations. The second possibility is that learners *do* attend to repetitions, but accord them little informational value when deriving inferences.

Determining which mechanism learners use will constrain how we develop a formal model of induction that seeks to accommodate both the tightening effect of types and the null effect of token presentations. That is, should a computational model accommodate the null effect of tokens as an encoding effect, probabilistically "dropping" some tokens from memory, or should we seek to do so as an inferential process, retaining all tokens but "downgrading" their evidentiary values?

Both approaches have precedent in the literature, and are sometimes assumed to

be interchangeable. For example, Navarro et al. (2012) introduced a variation on the original Bayesian inductive generalization model (Tenenbaum & Griffiths, 2001) that includes a "weighting" parameter θ that can be interpreted as a form of evidentiary downgrading. Yet the first case study (Tauber, Navarro, Perfors & Steyvers, 2017) proposed a Bayesian model of an inductive problem that included a "dropping" mechanism that relied on a probabilistic encoding parameter – also labelled θ – and incorrectly justified it by reference to the θ parameter in Navarro et al. (2012). This confusion is somewhat understandable, insofar as the two approaches yield almost identical predictions about the induction task, but they are not equivalent *in general* and it is a mistake to conflate them.

To determine which modeling approach we should proceed with, we designed a final experiment to tease apart the two competing explanations. Since both the "ignore" and "discount" processes can be expected to lead to null effects in generalization, we instead compare effects in memory performance. Participants observed data from the *1*, *1111*, or *4* conditions, then completed a surprise recognition memory test. Participants were asked to judge whether a training item and several novel items had been previously seen. If learners ignore repeated presentations, then additional presentations of a stimulus during training should not aid memory. That is, participants in the *4* condition should show similar recognition performance to participants in the *1* and *1111* conditions. On the other hand, if learners attend to repeated token presentations, recognition performance in the *4* condition should be superior to recognition in the other two conditions.

5.1. Method

5.1.1. Participants. Participants were 300 adults recruited from MTurk. Qualification and exclusion criteria were the same as in previous experiments. Data was excluded for two participants who failed the attention check question. The final sample size was 298 ($n = 123$ female, $n = 173$ male, $n = 1$ other, median age = 36 years). Participants were paid US$1.20 for the seven-minute task.

5.1.2. Design and Procedure. Participants were randomly assigned to either the *1*, *1111*, or *4* conditions in a between-subjects design. The training phase was identical to the "birds" trial of Experiment 1. As per that study, repeated token items were always presented in the same orientation. Unlike the earlier study however, we inserted a surprise recognition test between the training phase and the generalization test. For the recognition test, participants were instructed to "...indicate if you have already seen any of the following birds in the same orientation (that is, facing the same direction)". Each stimulus from the set of 18 memory items (described below) was then presented separately, in randomized order, and participants pressed an on-screen "Yes" or "No" button to indicate whether they had previously seen this bird facing this direction. Each memory item was visible until the participant made a response.

Performance was calculated on the basis of responses to 18 memory test items (all are included in Fig. S2 at https://osf.io/mtbve/). The test set included a single target item, which was the training phase stimulus presented in the *1* condition and common to the other conditions. The remaining 17 lures consisted of 1) the target item reflected horizontally, 2) all four high- and medium-similarity test stimuli from previous experiments, and 3) 12 other bird images that had not been used in previous experiments, but which the authors deemed were relatively similar to the target item. In pilot testing with an independent MTurk sample ($n = 60$), there was sufficient variability in false alarm rates to suggest that this set of lures was appropriate. The dependent variable was therefore recognition accuracy (hit rates - false alarm rates). Following the memory task, participants were instructed to "...think back to the original sample of 1-4 birds..." and the generalization test (with the seven generalization stimuli used in Experiment 1) proceeded as per normal.

5.2. Results and Discussion

Bayesian t-tests revealed very strong evidence for the alternative hypothesis that recognition accuracy was greater in the *4* condition ($M = .87$, $SD = .27$), compared to *1111* ($M = .62$, $SD = .37$; $BF_{10} > 1000$), and *1* ($M = .61$, $SD = .45$; $BF_{10} > 1000$).

Comparing hit rates only also supports the conclusion of greater recognition accuracy in the *4* condition (100%), compared to *1111* (90%; $BF_{10} > 1000$), and *1* (88%; $BF_{10} > 1000$). This suggests that participants did indeed attend to presentations, as the repeated presentations of *4* improved memory performance relative to *1* or *1111*.

This implies that the null effects of token presentations in previous experiments result from participants firstly attending to tokens, and then giving these tokens less weight than types when making generalization decisions. This result therefore provides evidence against the "ignore" hypothesis. We note that this experiment did not test a specific account of how repeated evidence is discounted - but it does show that such evidence is available in memory when inferences are made. In the next section, we formally implement a "discounting" process in a model of inductive generalization.

The results from the generalization test are described in the Supplementary Materials (https://osf.io/mtbve/). In general terms, we replicated the effects previously found – there was a tightening effect from *1* compared to *1111* ($BF_{10} = 23.00$ for main effect of types), and a null effect on generalization ratings as token presentations increased from *1* to *4* ($BF_{01} = 5.60$ for main effect of tokens).

6. Modeling

At the beginning of the paper, we motivated the work in part by referring to the predictions of computational models of inductive reasoning. For example, in Fig. 2 we provided an illustration of how the Bayesian framework introduced by Tenenbaum and Griffiths (2001) makes the correct prediction about the effect of adding new types. We also noted in Section 5 that there are existing variations of this model (e.g. Navarro et al., 2012) that could – at least in principle – explain the differences between how people generalize from types and tokens. Intuitively, the results appear to be in accordance with existing theoretical models, and we present modelling that suggests this intuition is correct.

In this section, we examine how Bayesian models of induction can be extended to account for 1) the tightening of property generalization when types are added to an

evidence sample, and 2) the absence of an effect of token presentations on generalization. Importantly, our goal in this section is to simulate only the *qualitative* pattern of results. The underlying stimulus materials (birds and flowers) are complex objects that are linked to rich semantic representations, and any formalism we apply to our data are necessarily gross simplifications.[2] With this in mind our approach is to take the model from Navarro et al. (2012) as our starting point, fix the parameters that describe the "stimulus representation" in the similarity space at intuitively plausible values, and then show that the pattern of results in the empirical data can be accommodated naturally by assuming that the weighting parameter θ differs when new data arrive in the form of additional types rather than additional tokens.

6.1. Specifying the Bayesian framework

The model described by Navarro et al. (2012) presents a Bayesian account of inductive generalization suitable for stimuli that vary along one relevant dimension[3]. Following Shepard (1987) the model assumes that the generalization of the novel property (e.g., gabbro bones) occupies a connected region along the relevant stimulus dimension. The set of all possible regions is assumed to be the *hypothesis space* $\mathcal{H}$ for the inductive problem. Following the explicitly Bayesian formulation introduced by

---

[2] Two examples of why this may be instructive. Firstly, we use a one-dimensional similarity space to represent the stimuli used in each experiment. However, we used complex multidimensional stimuli in our experiments — therefore the exact similarity space values, while interpretable in a relative sense (e.g., high-similarity stimuli are 'closer' to the training sample than medium-similarity stimuli), are ultimately arbitrary (high-similarity stimuli are not necessarily 0.2 units closer than medium-similarity stimuli). Secondly, we make the simplifying assumption that participants' generalization beliefs map linearly onto the 1-10 rating scale used. A more sophisticated modelling approach that estimates the mapping function can be found in the Appendix to the original Navarro et al. (2012) paper, along with derivations of the model predictions in closed form, but in that paper the stimuli were sufficiently simple that this kind of modelling was possible. We do not believe it is sensible to do so for the stimuli in these experiments.

[3] More precisely, when the relevant variation among items can be represented as such, as is the case when test items can be ordered along a continuum of similarity to the training items.

Tenenbaum and Griffiths (2001), the posterior probability $P(h|x)$ that any specific hypothesis $h \in \mathcal{H}$ describes the true generalization of the novel property given that the learner has observed the data x provided in the training set is given by Bayes rule

$$P(h \mid x) = \frac{P(x \mid h)\, P(h)}{\int_{h' \in \mathcal{H}} P(x \mid h')\ P(h')\ dh'} \tag{1}$$

where $P(h)$ specifies the prior plausibility of hypothesis h and the likelihood $P(x|h)$ describes the probability with which data x would be observed – according to the learner's subjective model of the world – if h were true. The probability of generalizing from training data x to test item y is obtained by integrating over hypotheses that include y in the generalization of the property. This set is denoted $h|y \in h$, and thus the generalization probability is

$$P(y \in h \mid x) = \int_{h|y \in h} P(h \mid x)\ dh \tag{2}$$

Closed form solutions to this integral are derived in the Appendix to Navarro et al. (2012), under a variety of priors $P(h)$ and likelihoods $P(x|h)$. For the purposes of this paper we assume the prior is uniform across the hypothesis space $\mathcal{H}$, and that the likelihood can be expressed in the following form:[4]

$$P(x \mid h, \theta) = \begin{cases} (1-\theta)\frac{1}{|\mathcal{X}|} + \theta\frac{1}{|h|}, & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $|\mathcal{X}|$ represents the *size* of the stimulus space $\mathcal{X}$, which in this case would correspond to the difference between the upper and lower bounds on similarity, and $|h|$ represents the size of a specific hypothesis h (i.e., how much of the sample space $\mathcal{X}$ is presumed to possess the novel property).

[4] Note that the notation in the current paper is more informal than the original paper. We take x to denote all information available in the training set, whereas Navarro et al. (2012) were more precise in separating the perceptual characteristics of a stimulus from the information specifying the novel property that it possesses. Accordingly, the expressions in our paper are more concise than those in the original.

6.2. Interpreting the model

Setting the formal details aside, the important psychological property of the model is specified using the weighting parameter θ. When $\theta = 0$ the learner is presumed to rely on a "weak sampling" assumption (per Shepard, 1987). Under this weak sampling assumption, when the learner encounters a new entity known to possess the novel property, the *only* information that it conveys to them is that fact: being told that "parrots have gabbro bones" will act to falsify any hypotheses h that do not allow parrots to have gabbro bones, but will have no *other* effect on the learner's beliefs. Under this assumption, repeated presentations have no effect whatsoever on the learner's inferences. If I already believe that parrots have gabbro bones, then I will encode repetitions of that fact (viz. Experiment 4), but these repetitions carry no weight for evaluating inductive hypotheses.

At the other end of the spectrum, setting $\theta = 1$ yields a "strong sampling" assumption (per Tenenbaum & Griffiths, 2001). Under the strong sampling model, the observed training item is presumed to be sampled from the set of all entities that fall within the generalization of h, and – importantly – that this sampling is independent of any previous samples. Under a strong sampling assumption, repeated presentations are very informative. If I sample 20 animals at random from the category of "things that have gabbro bones" and they all turn out to be parrots, I have much stronger evidence that parrots are the only animal that have gabbro bones than if I had only sampled a single parrot. Under such an account the effect of increased token presentations is to cause people to generalize more narrowly.

Taken together, these observations allow us to stipulate that the value of θ should be small when the training item is a true repetition: under such cases the item is *not* a new sample from the hypothesis h, it is a copy of an old observation x and something akin to weak sampling should apply. Note that we are not strictly equating tokens with weak sampling as described by Tenenbaum and Griffiths (2001) and Navarro et al. (2012) – rather, in our approach, a small θ value represents an understanding that observations are statistically dependent, so that later observations may contain the

same information as earlier observations (we return to the issue of modeling different sampling assumptions in the General Discussion Section 7.2.). In contrast, when a genuinely new type is observed people should recognize that this observation is informative (even if it looks indistinguishable from a previous one), and something more like the strong sampling model should apply. With this in mind, all subsequent simulations fix $\theta = .3$ when types are generated, and $\theta = .15$ when token presentations occur. The θ value for types is comparable to the empirical estimates reported by Ransom et al. (2016) in a property induction paradigm. By using these θ values, we could reproduce most of the ordinal patterns observed in our data, as the following sections outline.

6.3. Experiment 1

To model the results from Experiment 1 we ran the model 1000 times, each time setting different values for the ancillary parameters and averaging the results. The stimulus similarity space $\mathcal{X}$ is presumed (somewhat arbitrarily) to span a range from .4 to .9 along some relevant dimension: the training instances were located at a value of approximately .45, but could vary from one simulation to the next. More precisely, locations of new types were sampled from normal distributions with distinct means $(.45, .46, .44, .47)$ and a common standard deviation (.06). By contrast, the similarity space value of tokens was drawn from a normal distribution centered on the same value as the first type (.45) but with a narrower distribution ($SD = .009$), reflecting the fact that additional tokens were exact copies of one another (and so any variation in stimuli is purely perceptual noise), whereas new types represented discriminably different images and could thus be more variable.

Fig. 8 shows the results of the simulations. The similarity space values of four types and four tokens are represented by the filled circles in the upper-left corners of all panels in Fig. 8. The filled circles show that types (Fig. 8A-C) occupy a broader similarity space than tokens (Fig. 8D-F). The three dashed vertical lines in Fig. 8 represent the high-, medium-, and low-similarity test stimuli. Similarity is inversely

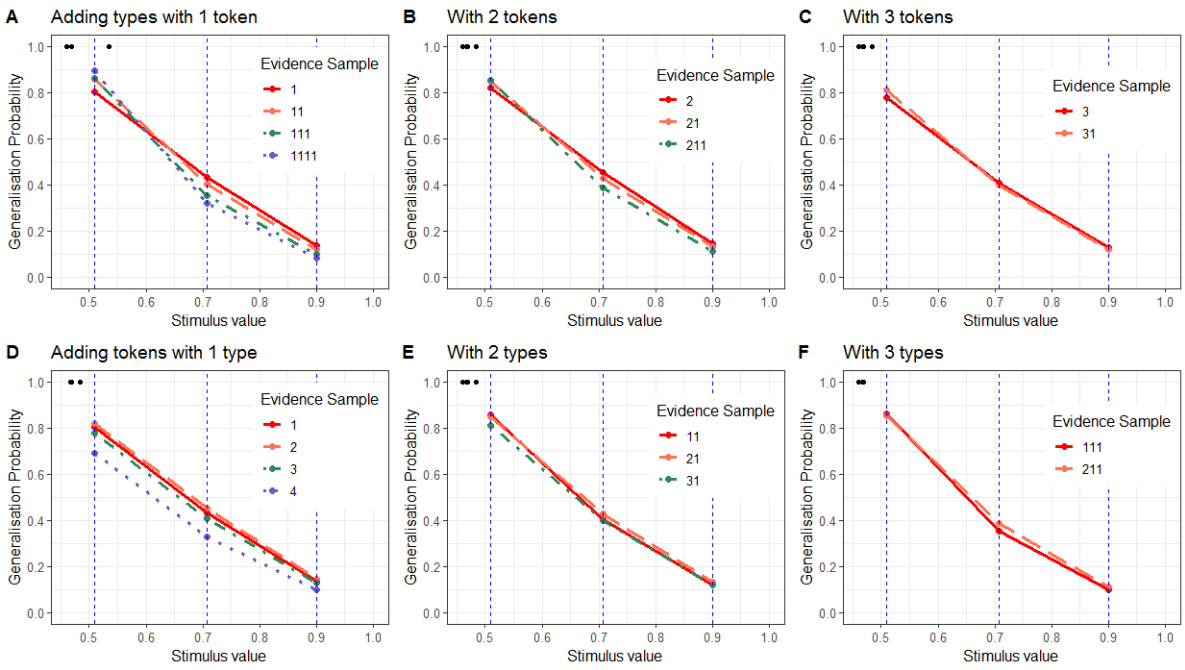

Figure 8. Modeling Experiment 1: The filled circles represent types or tokens, and the dashed vertical lines represent the high-, medium-, and low-similarity test stimuli. The Bayesian model of inductive reasoning predicts tightening as the number of types increases (Panels A-C) and decreasing, but largely similar, generalization as the number of tokens increase (Panels D-F). These modeling results can be compared with participant ratings shown in Fig. 5B-D and F-H.

related to distance from the evidence sample; thus, the high-similarity test stimuli are 'closest' to the evidence sample (similarity space value = .51), the low-similarity test stimuli are furthest away (similarity space value = .9), and the medium-similarity test stimuli are roughly equidistant between the high- and low-similarity categories (similarity space value = .7).

Fig. 8A represents generalization from the evidence samples *1*, *11*, *111*, and *1111* (i.e., one, two, three, and four types), and can therefore be compared with the participant generalization ratings of Fig. 5B. The model captures the observed pattern that adding types increased generalization to high-similarity test items, and decreased generalization to medium- and low-similarity items. This is tightening, consistent with participant performance in Experiment 1. Similarly, it produced weaker tightening

effects (i.e., smaller difference between evidence samples) when adding types with two and three token presentations (Fig. 8B-C), which is also consistent with participant generalization ratings (Fig. 5C-D).

Fig. 8D represents generalization from the evidence samples *1*, *2*, *3*, and *4*. Additional presentations decreased generalization ratings across all test categories, however these differences are very small. This is comparable to the participant ratings observed in Experiment 1 (Fig. 5F). Null effects predicted by the model with two and three types (Fig. 8E-F) are also consistent with participant generalization ratings (Fig. 5G-H).

The overall pattern of results suggests that a version of the Navarro et al. (2012) model can reproduce the *qualitative* phenomena of interest, but it does not always capture the magnitudes of the various effects correctly (also see the Supplementary Materials for scatterplots comparing mean participant ratings with model predictions in this and other experiments).

6.4. Experiment 2

The parameters used to simulate the types of Experiment 1 are again used to simulate Experiment 2 (except that Experiment 2 only had three types, thus only the first three similarity space values used in Experiment 1 are used here). For tokens, the standard deviation is increased slightly from .009 in Experiment 1 to .01 in Experiment 2, because the images used to represent tokens in Experiment 1 were rotated and/or reflected in Experiment 2, and therefore more dissimilar to each other. The results of this simulation are shown in Fig. 9. The effect of adding types that are each presented twice is shown in Panel A, while the effect of adding types that are each presented four times is shown in Panel B. In both panels, adding types produced tightening, matching the ordinal pattern observed in participant data for medium- and low-similarity test items. For high-similarity test items however, the model predicted that generalization would be lowest when only one type has been observed in the evidence sample. However, as discussed in Section 3.2., there is no evidence for that effect in the

participant data with high-similarity stimuli. Overall then, the model produced behavior that mostly accords with our intuitions and with empirical data but is not fully supported by participants' ratings for high-similarity stimuli.

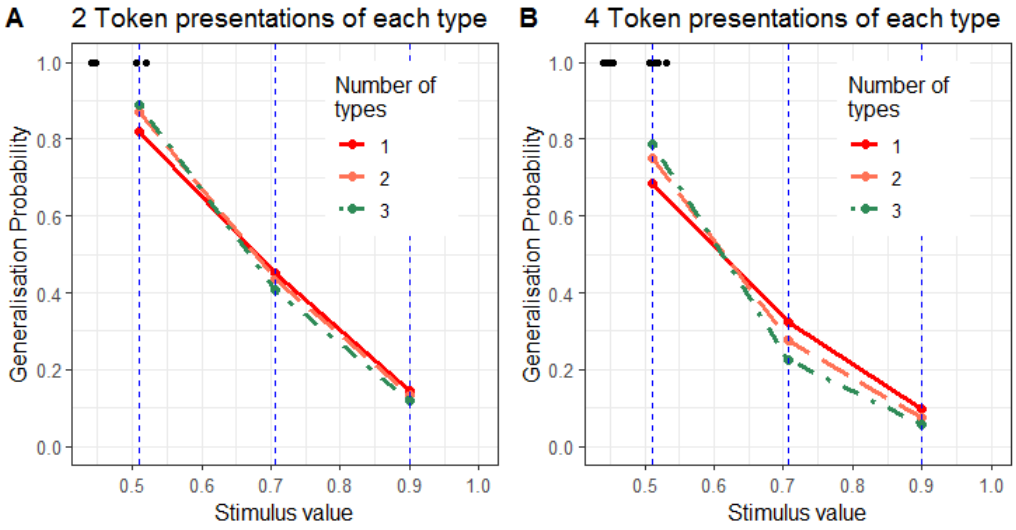

Figure 9. Modeling Experiment 2: The Bayesian model of inductive reasoning predicts a tightening effect as types are added. This occurs when each type is repeated with A) two and B) four token presentations.

6.5. Experiment 3

Experiment 3 differed from Experiments 1 and 2 because the 'types' used in this study were represented by copies of the one image. Hence, we changed two parameters to reflect the fact that the type-label instances in Experiment 3 were more similar to each other than the types used in Experiments 1 and 2. Firstly, we defined six new means to center the distributions of the six type-label instances. These new means (.452, .448, .44, .444, .456, .46) were within a narrower range than those used in Experiments 1 and 2. Secondly, the standard deviation for the distributions representing type-label instances ($SD = .03$) was half that used in Experiments 1 and 2. The parameters for token-label instances were the same as previously used for tokens in Experiment 2, since the stimuli and cover stories were unchanged ($SD = .01$, $\theta = .15$). The theta value for types-label instances was also the same value used for types in Experiment 2 ($\theta = .3$).

The model behavior is shown in Fig. 10. Recall that participants' generalization ratings decreased with an increasing number of type-label instances, but were largely unchanged with an increasing number of token-label instances. Fig. 10 shows that the model did not capture this difference. Instead, the model predicted decreasing generalization ratings for both type-label and token-label instances. The misfit is evident in the very large difference between conditions with an increasing number of token-label instances (Fig. 10B).

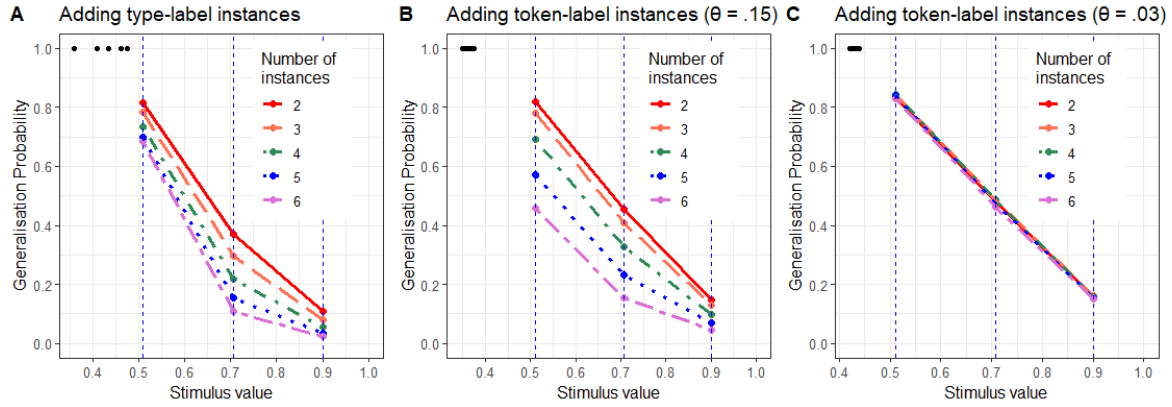

Figure 10. Modeling Experiment 3: The Bayesian model of inductive reasoning predicts that A) adding type-label instances decreases generalization ratings, as does B) adding token-label instances. C) Adding token-label instances with a smaller θ value more closely resembled the null effect observed in participant ratings.

As an exploratory venture, we tested how varying θ values changes model simulations for the token-label conditions. Lower θ values, implying less informational value of additional instances, decreased differences between conditions, and therefore resulted in better qualitative fit with participant data. As shown in Fig. 10C, a small θ value of .03 produced model predictions that resembled the null effect observed in participant ratings. This therefore suggests that a lower θ value than previously assumed for Experiments 1 and 2 is required to capture participants' generalization in Experiment 3.

This may be because Experiment 3 used a larger number of identical tokens (six) than Experiments 1 or 2 (four). As the number of token presentations increases, this

may reinforce the belief that the sampling process simply involves repetition of a single item. Six repeated presentations seems very unlikely under any other sampling process – and less likely than four repeated presentations. Therefore a smaller theta value – representing greater statistical dependence between observations, and thus less informational value – better captures participants' inferences in Experiment 3.

6.6. Model Robustness and Generality

In previous sections, we described qualitative model predictions that resulted from a specific combination of parameter values. Although we have justified each parameter choice in the context of experimental features, we could have plausibly chosen other values for each parameter. Therefore, we tested the robustness of our model predictions by examining how frequently the reported qualitative trends were reproduced with different parameter values. We ran 3000 simulations of the models used in Experiments 1-3: in each simulation, we re-sampled each parameter value (θ for types and tokens, mean and standard deviation values for sampling similarity space values for types and tokens, and similarity space values for the three test-training similarity levels) from plausible ranges around the parameter values reported in the previous sections. See Supplementary Materials for details of parameter ranges and simulation results; R code for the simulations can be found at https://osf.io/mtbve/.

We evaluated 30 patterns – each qualitative trend was evaluated separately at each level of test-training similarity (high-, medium-, low-similarity), and also for each number of types/tokens (e.g., for Experiment 1, the effect of adding types was evaluated separately for conditions with 1, 2, or 3 token presentations). Of these 30 patterns, 28 were reproduced at rates above chance across the simulated parameter space. The two exceptions (both involving predictions of tightening with additional types) are discussed in the Supplementary Materials. Combining reproduction rates across levels of test-training similarity and number of types/tokens, we found that: a) adding types caused tightening in 67.2% of Experiment 1 simulations and 72.0% of Experiment 2 simulations, b) adding tokens either decreased or did not change generalization in 96.8%

of Experiment 1 simulations, and c) adding both type-label and token-label instances decreased generalization in 84.1% of Experiment 3 simulations. These results indicate that our main model predictions about the relative effects of adding types and tokens are quite general, and not dependent on the specific parameter values chosen for our original model implementation.

7. General Discussion

The aim of the present work was to compare the effects on inductive inference of adding unique types or repeated token presentations to a training sample. Although the effect of adding types has been studied, the effect of tokens has not. Given how frequently everyday samples of evidence contain repetitions, and a large body of evidence that suggests a substantial effect of repetitions (e.g., Barsalou et al., 1998; Nosofsky, 1988; Unkelbach, Fiedler & Freytag, 2007), we sought to identify the effect of token presentations in inductive inference.

7.1. Types produce tightening while tokens are discounted

Tightening, in which additional types or unique instances of a particular category decrease generalization to stimuli outside of that category (and may increase generalization to stimuli within that category), is consistent with many existing Bayesian models of induction (e.g., Navarro et al., 2012; Tenenbaum & Griffiths, 2001). By assuming that types in the evidence sample have been strongly sampled, Bayesian models predict that observing more types within a narrow stimulus space should cause the learner to prefer smaller hypotheses over larger hypotheses. The tightening that we observed in Experiments 1 and 2 therefore provides further support for the importance of sampling assumptions in property induction and further evidence for the size principle.

Notably, the current work goes beyond previous demonstrations of tightening. We provided the first evidence that the same sample of multidimensional, biological stimuli can increase property generalization to near stimuli and decrease generalization to more distant stimuli. In addition, we found that this tightening effect persisted when types

were presented multiple times (Experiments 1 and 2). Adding types also constrained generalization when types were portrayed by visually-identical stimuli (Experiment 3).

This was the first study to explicitly examine the role of token presentations in property inference. Experiments 1 to 3 consistently found evidence for the null hypothesis that repeated presentations did not affect property generalization. In Experiment 3, we showed that this null effect did not hold when repeated images were used to represent unique types, thus highlighting the importance of learners' beliefs about how instances are sampled. Experiment 4 suggested that the null effect of token presentations occurs even though participants encode them in memory, implying that participants discount the informational value of tokens during the inference process. Overall, we found strong evidence that participants discriminate between the informational value of additional types and tokens when making property inferences.

7.2. Implications for models of induction

Until now, models of induction have not explicitly attempted to implement the effect of token presentations. By connecting stronger sampling assumptions with types and weaker sampling assumptions with tokens, our adaptation of Navarro et al.'s (2012) model represents how we think participants make inferences: namely, by assigning greater informational value to types, and close-to-null informational value to tokens. We note that our use of θ differs from its typical instantiation. In previous Bayesian accounts, a small θ value reflected the learner's belief that instances are sampled randomly from a population of objects that may or may not have the property of interest (compared to an instance that is sampled only from the population of objects that do have the property). In our case, the θ value reflects the learner's understanding of the statistical dependence between observations. Small θ values reflect a belief that subsequent tokens repeat information that has previously been presented. To link this back to sampling processes, we could think of token presentations as arising from an extreme version of sampling with replacement (i.e., previously presented items are returned to a very small population pool and are likely to be resampled). Therefore,

repeatedly sampling one instance does not provide novel information about the population they are being sampled from, and thus repeated token presentations have little effect on the learner's inductive inferences.

Although Bayesian models have an advantage when considering such sampling effects (see also Hayes, Banner et al., 2019; Vong, Hendrickson, Perfors & Navarro, 2013; Voorspoels et al., 2015), we note that earlier models of induction can account for some of our empirical findings. Notably, the tightening effect could be explained by similarity-based models like the similarity-coverage model (SimCov; Osherson et al., 1990).

Under SimCov, the likelihood of generalizing is based on mean similarity 1) between premise and conclusion categories, and 2) between premise categories and the lowest-level superordinate category that encompasses both premise and conclusion categories. In our case, this would be the mean similarity 1) between the evidence sample and the generalization stimuli (other green parrots, other parrots, or other birds), and 2) between the evidence sample and the categories of 'green parrots', 'parrots', or 'birds', respectively. Using these similarity calculations, the similarity-coverage model can predict the tightening effect of adding types. Thinking firstly of the high-similarity stimuli (other green parrots), additional green parrots in the evidence sample serve to increase the mean similarity between these green parrots and the superordinate category of 'green parrots'. Thus, we would observe increased generalization ratings. However, when we think of the medium- and low-similarity stimuli, we can see that additional green parrots in the evidence sample now decrease the mean similarity between these green parrots and the broader superordinate categories of 'parrots' and 'birds'. This is because these larger categories involve more instances beyond the 'green parrot' subcategory. Therefore, observing additional instances that represent only a small subsection of that large category should decrease generalization ratings.

The framing effect of Experiment 3 is problematic for both SimCov and our current instantiation of a Bayesian model. Because SimCov does not consider

participant beliefs about the generative process behind evidence samples, it has no explanatory mechanism for the difference in generalization patterns between the type-label and token-label conditions of Experiment 3. Theoretically, the Bayesian model of induction specified by Navarro et al. (2012), which allows for different sampling assumptions, could. However, our initial parameter values did not reproduce this difference well. We found that only very small θ values (about one fifth the $\theta = .15$ initially assumed) produced the null effect of adding token-label instances. We speculated that smaller θ values may be required here because evidence samples in Experiment 3 contained more tokens, but further investigation should inform any other parameter changes. Despite the model's limited ability to simulate this subtle experimental manipulation, the current model *can* account for the gross effects of adding types compared to adding tokens using the same framework. This therefore extends the explanatory breadth of Bayesian models of induction, and shows that we need not assume that all items of evidence are novel or independent of other items.

7.3. Implications for how we understand repetitions

The current studies demonstrate that learners are sensitive to the different evidentiary value of new, as compared to repeated observations, in an inference task. This sensitivity was demonstrated even when the items presented as new or old observations were identical but subject to different cover stories about their origin (Experiment 3). There were no explicit instructions to discount repeated presentations, but Experiment 4 suggests that participants inferred the (lack of) informational value of tokens, and made the necessary correction by discarding them when generalizing to less-similar stimuli. These results are broadly consistent with Perfors et al.'s (2014) findings that instance repetition did not affect generalization decisions in a single category learning task. Such results demonstrate that learners can attend to repeated observations but subsequently treat these repetitions as having little evidentiary value.

This is contrary to Nosofsky's (1988) finding in which presentation frequency *did* affect typicality judgements and categorization decisions. Although categorization tasks

certainly involve inductive generalization, there are potentially important differences between inductive inference based on observations from a single category (as in the current experiments) and categorization decisions that involve two or more categories. Hendrickson et al. (2019) suggest that tasks involving samples drawn from only one category (as in most property induction studies) imply a strong sampling assumption, while tasks involving items from multiple categories often imply a weaker form of sampling. Different sampling assumptions could explain why increasing sample size (i.e., adding types) led to tightening in a one-category task, but not in a two-category task (Hendrickson et al., 2019). We speculate that this may also account for different token effects across induction and categorization tasks.

Beyond Nosofsky (1988), our findings are also broadly at odds with the numerous demonstrations that learners typically struggle to understand the implications of different types of evidence sampling when making judgments under uncertainty (Foster et al., 2012; Gonzalez, 1994; Nosofsky, 1988; Yousif et al., 2019). One popular account, termed "metacognitive myopia" specifically suggests that people often fail to take account of the different weight that should be attached to independent (i.e., novel) and dependent (redundant or repeated) evidence (Fiedler, 2012; Fiedler, Joschahofferbert, Krueger & Koch, 2015). For example, Fiedler, Hütter, Schott and Kutzner (2019) argue that "Metacognitive myopia is evident in the inability not to be influenced by selectively repeated arguments in preference learning. . . or in collective decision making. . . " (p. 2) — our findings suggest this 'inability' does not extend to property induction.

Why is it then that we find in favor of learners' ability to appropriately discount repeated presentations of old evidence, when so much other work has arrived at more pessimistic conclusions? We propose two overarching reasons: using a property induction task promotes attention to the quality of evidence, and our property induction task contained favorable task characteristics that facilitated discounting of old evidence. On the first point, we note that no other studies have specifically used a property induction paradigm. For example, Nosofsky (1988) used a color categorization task where, arguably, attending to the details of the stimuli themselves is more

important for the categorization decision than attending to the sampling process producing the stimuli. Conversely, in a property induction task, the goal is to generalize from a limited sample out to a broader population – in this context, it is important to assess *how* that sample was produced.

Other benefits to learners in our property induction task was the explicit labelling of old information, and the manageable cognitive load throughout. Our findings in Experiment 3 showed that learners' beliefs about whether or not an individual datum is new or old is critical. In many ways, our experiments provided the ideal conditions for this discrimination. Repeated items were clearly labelled as such (through repeated ID codes), and the accompanying cover story supported this interpretation. The current experiments also minimized demands on memory – participants could directly inspect each additional instance and compare it to previous instances. This made it easy for the learner to discriminate between new and repeated evidence, and assign the appropriate informational value to each instance.

As mentioned earlier, many previous experiments – and the real world – often involved less "kind" learning environments (e.g., Jasny et al., 2015; Nosofsky, 1988; Unkelbach et al., 2007; Yousif et al., 2019). In real-world environments, the dependency amongst sources can be obscured. For example, in an analysis of science-denier blogs, Harvey et al. (2018) found that about 80% of such blogs relied on one particular blog as the primary source when debating the impacts of climate change on polar bears. Without attending to this source dependence, the sheer quantity of such denier blogs can be convincing.

In previous experiments, the memory load required for tracking the number of types or tokens was typically greater than our maximum evidence sample size of 12. In Unkelbach et al. (2007), participants monitored 10 shares over 16 time points ('trading days'), Barsalou et al. (1998) presented 30 exemplars of five different fish, while Nosofsky (1988) presented 48 instances of 12 colours. Moreover, these previous experiments presented repeated items sequentially so that identification of repeats required monitoring of memory contents. It is likely that adding such memory demands

reduces the learner's ability to discount repeated information. Supporting this, Enke and Zimmermann (2017) found that 'correlation neglect', or the failure to correct for a dependency between information sources, increased as the amount of information (i.e., number of messages in the form of computer signals) increased.

The current results suggest that when learning conditions are conducive to easy discrimination between old and new information and memory demands are minimized, learners do differentiate in the weight they attach to old and new evidence. A challenge for future research is to identify the extent to which this ability is preserved as these conditions are relaxed and the learning environment becomes more complex.

8. Conclusions

When making inductive inferences outside the laboratory, learners are likely to encounter a variety of evidence. Evidence may take the form of new independent facts that we learn (additional "types" in the current experiments) or repetitions of previously learned facts (additional "token" presentations). The current work is the first to show that learners treat these different forms of evidence in very different ways when doing induction. New independent evidence constrains inferences about how far a property generalizes (e.g., via tightening). By contrast, repetitions of old evidence are certainly encoded, but are given little weight when making inferences. A Bayesian model of induction that assigns less informational value to tokens as compared to types can capture both of these effects, and hence seems like a promising framework for the future study of the effect of repeated evidence on human inference.

References

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, *26*(10), 1531–1542. doi:10.1177/0956797615594620

Barsalou, L. W., Huttenlocher, J. & Lamberts, K. (1998). Basing categorization on individuals and events. *Cognitive Psychology*, *36*(3), 203–272. doi:10.1006/cogp.1998.0687

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E. & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330. doi:10.1016/j.cognition.2010.10.001

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. doi:10.3758/s13428-014-0458-y

Enke, B. & Zimmermann, F. (2017). Correlation Neglect in Belief Formation. *The Review of Economic Studies*, *86*(1), 313–332. doi:10.1093/restud/rdx081

Feeney, A. (2007). How many processes underlie category-based induction? Effects of conclusion specificity and cognitive ability. *Memory & Cognition*, *35*(7), 1830–1839. doi:10.3758/BF03193513

Feeney, A. & Heit, E. (Eds.). (2007). *Inductive reasoning: Experimental, developmental, and computational approaches*. New York, NY: Cambridge University Press.

Fiedler, K. (2012). Meta-Cognitive Myopia and the Dilemmas of Inductive-Statistical Inference. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 57, pp. 1–55). doi:10.1016/B978-0-12-394293-7.00001-7

Fiedler, K., Hütter, M., Schott, M. & Kutzner, F. (2019). Metacognitive myopia and the overutilization of misleading advice. *Journal of Behavioral Decision Making*, *32*(3), 317–333. doi:10.1002/bdm.2109

Fiedler, K., Joschahofferbert, F., Krueger, T. & Koch, A. (2015). The Tragedy of Democratic Decision Making. In J. P. Forgas, K. Fiedler & W. D. Crano (Eds.),

Social Psychology and Politics (Vol. 17, pp. 193–208). New York, NY: Psychology Press.

Foster, J. L., Huthwaite, T., Yesberg, J. A., Garry, M. & Loftus, E. F. (2012). Repetition, not number of sources, increases both susceptibility to misinformation and confidence in the accuracy of eyewitnesses. *Acta psychologica*, *139*(2), 320–326. doi:10.1016/j.actpsy.2011.12.004

Gonzalez, R. (1994). When Words Speak Louder Than Actions: Another′s Evaluations Can Appear More Diagnostic Than Their Decisions. *Organizational Behavior and Human Decision Processes*, *58*(2), 214–245. doi:10.1006/obhd.1994.1035

Gutheil, G. & Gelman, S. A. (1997). Children's Use of Sample Size and Diversity Information within Basic-Level Categories. *Journal of Experimental Child Psychology*, *64*(2), 159–174. doi:10.1006/jecp.1996.2344

Gvirsman, S. D. (2014). It's Not That We Don't Know, It's That We Don't Care: Explaining Why Selective Exposure Polarizes Attitudes. *Mass Communication & Society*, *17*(1), 74–97. doi:10.1080/15205436.2013.816738

Hahn, U., von Sydow, M. & Merdes, C. (2019). How Communication Can Make Voters Choose Less Well. *Topics in Cognitive Science*, *11*(1), 194–206. doi:10.1111/tops.12401

Harris, A. J. L., Hahn, U., Madsen, J. K. & Hsu, A. S. (2016). The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. *Cognitive Science*, *40*(6), 1496–1533. doi:10.1111/cogs.12276

Harvey, J. A., van den Berg, D., Ellers, J., Kampen, R., Crowther, T. W., Roessingh, P., ... Mann, M. E. (2018). Internet Blogs, Polar Bears, and Climate-Change Denial by Proxy. *BioScience*, *68*(4), 281–287. doi:10.1093/biosci/bix133

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. doi:10.1016/S0022-5371(77)80012-1

Hayes, B. K., Banner, S., Forrester, S. & Navarro, D. J. (2019). Selective sampling and inductive inference: Drawing inferences based on observed and missing evidence. *Cognitive Psychology*, *113*, 101221. doi:10.1016/j.cogpsych.2019.05.003

Hayes, B. K. & Heit, E. (2017). Inductive reasoning 2.0. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(3), e1459. doi:10.1002/wcs.1459

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K. & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review*, *26*(3), 1043–1050. doi:10.3758/s13423-018-1562-2

Heit, E. & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 411–422. doi:10.1037/0278-7393.20.2.411

Hendrickson, A. T., Perfors, A., Navarro, D. J. & Ransom, K. (2019). Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Cognitive Psychology*, *111*, 80–102. doi:10.1016/j.cogpsych.2019.03.001

Iyengar, S. & Hahn, K. S. (2009). Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication*, *59*(1), 19–39. doi:10.1111/j.1460-2466.2008.01402.x

Jasny, L., Waggle, J. & Fisher, D. R. (2015). An empirical examination of echo chambers in US climate policy networks. *Nature Climate Change*, *5*(8), 782–786. doi:10.1038/nclimate2666

Kemp, C. & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58. doi:10.1037/a0014282

Lee, J. C., Lovibond, P. F., Hayes, B. K. & Navarro, D. J. (2019). Negative evidence and inductive reasoning in generalization of associative learning. *Journal of Experimental Psychology: General*, *148*(2), 289–303. doi:10.1037/xge0000496

Maines, L. A. (1996). An experimental examination of subjective forecast combination. *International Journal of Forecasting*, *12*(2), 223–233. doi:10.1016/0169-2070(95)00623-0

Medin, D. L., Coley, J. D., Storms, G. & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517–532. doi:10.3758/BF03196515

Morey, R. D. & Rouder, J. N. (2015). BayesFactor. R package version 0.9.12-2.

Navarro, D. J., Dry, M. J. & Lee, M. D. (2012). Sampling Assumptions in Inductive Generalization. *Cognitive Science*, *36*(2), 187–223. doi:10.1111/j.1551-6709.2011.01212.x

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. doi:10.1037/rev0000077

Navarro, D. J. & Perfors, A. F. (2010). Similarity, feature discovery, and the size principle. *Acta Psychologica*, *133*(3), 256–268. doi:10.1016/j.actpsy.2009.10.008

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65. doi:10.1037/0278-7393.14.1.54

Osherson, D. N., Smith, E. E., Wilkie, O., López, A. & Shafir, Eldar. (1990). Category-based induction. *Psychological Review*, *97*(2), 185–200. doi:10.1037/0033-295X.97.2.185

Perfors, A., Ransom, K. & Navarro, D. (2014). People ignore token frequency when deciding how widely to generalize. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (Vol. 36, pp. 2759–2764). Austin, TX: Cognitive Science Society.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111–163. doi:10.2307/271063

Ransom, K. J., Perfors, A. & Navarro, D. J. (2016). Leaping to Conclusions: Why Premise Relevance Affects Argument Strength. *Cognitive Science*, *40*(7), 1775–1796. doi:10.1111/cogs.12308

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. doi:10.1016/j.jmp.2012.08.001

Sanjana, N. E. & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. In S. Becker, S. Thrun & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems (NIPS)* (pp. 59–66). MIT Press. Retrieved from http://papers.nips.cc/paper/2284-bayesian-models-of-inductive-generalization.pdf

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A. & Menczer, F. (2019). On the Inevitability of Online Echo Chambers. *arXiv preprint arXiv:1905.03919*.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. doi:10.1126/science.3629243

Tauber, S., Navarro, D. J., Perfors, A. & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(4), 410–441. doi:10.1037/rev0000052

Tenenbaum, J. B. & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. doi:10.1017/S0140525X01000061

Unkelbach, C., Fiedler, K. & Freytag, P. (2007). Information repetition in evaluative judgments: Easy to monitor, hard to control. *Organizational Behavior and Human Decision Processes*, *103*(1), 37–52. doi:10.1016/j.obhdp.2006.12.002

Unkelbach, C. & Rom, S. C. (2017). A referential theory of the repetition-induced truth effect. *Cognition*, *160*, 110–126. doi:10.1016/j.cognition.2016.12.016

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 Meeting 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. doi:10.1016/j.cogpsych.2015.07.001

Whalen, A., Griffiths, T. L. & Buchsbaum, D. (2017). Sensitivity to Shared Information in Social Learning. *Cognitive Science*. doi:10.1111/cogs.12485

Xu, F. & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272. doi:10.1037/0033-295X.114.2.245

Yousif, S. R., Aboody, R. & Keil, F. C. (2019). The Illusion of Consensus: A Failure to Distinguish Between True and False Consensus. *Psychological Science*, *30*(8), 1195–1204. doi:10.1177/0956797619856844

Appendix A - Verbatim instructions for Experiment 3

Token-label instances:

When a bird has gabbro bones, your research assistants photograph and record the bird, giving each individual bird a unique ID number.

This ID number is written on a small tag attached to each bird's leg. This ensures that birds are not counted twice.

However, the research assistants work independently at different sites, so you **may see multiple photographs of the same bird**.

You will be able to tell if it is the **same bird** because the **ID number will be the same**.

Type-label instances:

When a bird has gabbro bones, your research assistants photograph and record the bird, giving each individual bird a unique ID number.

This ID number is recorded on the research assistant's photograph.

The research assistants work together at the same site, so the **same bird is never photographed more than once**.

You may see **similar-looking birds**, but they are different birds with **different ID numbers**.