How many observations is one generic worth?

Michael Henry Tessler (tessler@mit.edu)
Department of Brain and Cognitive Sciences, MIT

Sophie Bridgers (sbridge@stanford.edu)
Department of Psychology, Stanford University

Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences, MIT

Abstract

Generic language (e.g., "Birds fly") conveys generalizations about categories and is essential for learning beyond our direct experience. The meaning of generic language is notoriously hard to specify, however (e.g., penguins don't fly). Tessler and Goodman (2019b) proposed a model for generics that is mathematically equivalent to Bayesian belief-updating based on a single pedagogical example, suggesting a deep connection between learning from experience and learning from language. Relatedly, Csibra and Shamsudheen (2015) argue that generics are inherently pedagogical, understood by infants as referring to a member of a kind. In two experiments with adults, we quantify the exchange-rate between generics and observations by relating their belief-updating capacity, varying both the number of observations and whether they are presented pedagogically or incidentally. We find generics convey stronger generalizations than single pedagogical observations (Expt. 1), even when the property is explicitly demarcated (Expt. 2). We suggest revisions to the vague quantifier model of generics that would allow it to accommodate this intriguing exchange-rate.

Keywords: generic language; Bayesian learning; belief updating; pedagogical sampling; observational learning

Introduction

The world is a confusing and confounding place, but forming the right kinds of generalizations eases our navigation of the environment. One major route for acquiring generalizable knowledge is from observations. Indeed, one hallmark of human intelligence, present in infancy and childhood, is our capacity to draw strong generalizations from just a few examples (e.g., Gopnik, Sobel, Schulz, & Glymour, 2001; Gopnik et al., 2004; Gweon, Tenenbaum, & Schulz, 2010; Tenenbaum, Griffiths, & Kemp, 2006). At the same time, abstract generalizations can also be conveyed with language, using what is called *generic language* (or, *generics*; e.g., "Swans are white"; Carlson, 1977; Leslie, 2007; Gelman, Star, & Flukes, 2002; Tessler & Goodman, 2019a). Given that generalizations can be acquired from observation and from language, then there must be some relationship between the two. There must be some point at which the strength of an inductive generalization drawn from experience is equal to that of a generalization learned from language (Fig. 1).

Not all observations are created equal. Watching an informed and cooperative interlocutor intentionally convey an example via a demonstration is a stronger signal than if the observation is observed by happenstance (Shafto, Goodman, & Frank, 2012), which can result in more robust generalizations in adults and children (Goodman, Baker, & Tenenbaum, 2009; Butler & Markman, 2012). Thus, a crucial question is not only how many observations is one generic worth, but what kind of observations are they – socially demonstrated or just incidentally observed?

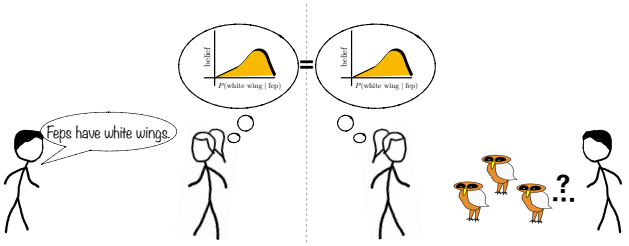

Figure 1: Generalizations about categories – expressed via degrees of belief that an instance of the category will have the feature – are learned both from generic language (left) and direct observations (right). When is the strength of generalization drawn from examples equal to that drawn from a generic statement? Or, how many observations is one generic worth?

The precise relationship between learning from examples vs. from language is difficult to articulate because learning from linguistic utterances operates via the *truth conditions* of the utterance, which are often difficult to specify precisely. Generics are a clear case of this squishiness: while "Triangles have three sides" should be taken to mean that exactly 100% of triangles have three sides, "Swans are white" is more tolerating of exceptions (i.e., there are black swans); "Mosquitoes carry malaria" is an example of a generic that conveys a very weak generalization: the vast majority of real-world mosquitoes do not carry the virus. To explain this heterogeneity, Tessler and Goodman (2019a) proposed a meaning for generics that is similar to that of quantifiers (e.g., *some*, *most*, or *all*) but which has an uncertain truth-conditional threshold; that is, the threshold beyond which the generic is literally true is underspecified but inferred in context. The generics model assumes a uniform prior distribution over thresholds (i.e., all values of the generic threshold are equally likely *a priori*). Tessler and Goodman (2019b) show that this assumption makes the model identical to rational Bayesian updating from a single positive observation (e.g., if trying to infer the weight of a coin, flipping the coin once,

and observing it land on heads) and extend the model so that the generalizations learned from generics can be strengthened through pragmatic reasoning, analogous to learning from a pedagogically sampled example (Shafto et al., 2012).

The relationship between the meaning of generics and pedagogical examples has independently been interrogated and elucidated to understand infant cognition. Csibra and Shamsudheen (2015) argue that when preverbal infants observe an instance of a novel category (call it a *blicket*), they not only have the capacity to individuate this object as a singular entity (i.e., *this is a blicket*) but also have the capacity to see the object as an index to the kind (i.e., this blicket is a pointer to the kind BLICKETS). Because of infants' sensitivity to ostensive cues (i.e., natural pedagogy; Csibra & Gergely, 2009), when an object is presented to an infant with pedagogical cues, the infant can interpret the object, not as a singular entity, but as an index to the kind; then, if a property is predicated of that object (e.g., the blicket is shown to squeak), it will be taken by the infant to apply to the kind as a sort of non-verbal generic: *Blickets squeak*.[1] This view thus also draws a direct connection between generics and a single, pedagogical example. [2]

Thus, proposals from two rather different theoretical frameworks—Bayesian models of semantics/pragmatics and infant cognition—point to the rather intriguing hypothesis that the information content of a generic is equivalent to that of a single, pedagogically-presented example. On the other hand, generics are commonly expressed with plurals in many languages including English (e.g., *Dogs bark*), and a plural should be a cue that the literal meaning goes beyond a single example. Furthermore, the relationship between generics and pedagogical examples that Csibra and Shamsudheen (2015) propose for preverbal infants may not be the same throughout development; indeed, 3- and 4-year-olds can interpret the ostensive cue of pointing as a signal that the information conveyed is not generalizable, but rather specific to the exemplars referenced by the point (Meyer & Baldwin, 2013).

In this paper, we take an empirical approach to investigate the relationship between learning from examples vs. generics by attempting to quantify the *exchange rate* between generics and observations. Contra the theoretical proposals, we find that in adults, a generic is worth at least two pedagogically sampled examples. We discuss the implications of this relationship and describe some of its boundary conditions.

Experiments

We develop an empirical paradigm where participants learn about a novel category from examples, from generic language, or both. Participants are then asked to judge the likelihood that a future instance of a category would have the property (cf. Gelman et al., 2002; Cimpian, Brandone, & Gelman, 2010; Tessler & Goodman, 2019b). We titrate the number of examples participants observe – as well as manipulate the communicative intent behind the observations – in order to determine the point at which the strength of the generalization implied by examples is equal to that of a generic statement (i.e., the *exchange rate* between generics and observations). Number of examples and communicative intent were manipulated between-participants; no participant completed multiple conditions. Experimental paradigms, data, models, and analysis scripts can be found at `github.com/mhtess/genex_cogsci2020`.

Experiment 1

Participants We recruited 465 adult participants from Amazon's Mechanical Turk. By experimenter error, 38 participants were able to complete the experiment multiple times (comprising a total of 106 submissions); we used each participant's first submission only, leaving 397 submissions from unique participants. Participants were restricted to those with U.S. IP addresses with at least a 95% work approval rating.

Materials We used exemplars from three semi-novel categories (bird, flower, artifact) labeled with novel labels (*fep*, *dax*, *blicket*). Each exemplar had a particular feature that was highlighted in the learning phase of the experiment: the color of the wing of the bird (a *white wing*), the color of the center of the flower (a *black center*), or the sound that the artifact produced (*squeaking*). We chose these somewhat atypical features so that it would be plausible the feature could be prevalent in varying degrees (e.g., the color of a bird's wing can vary by sub-species as well as by individuals) in order to increase the dynamic range of our dependent measure.

Procedure The experiment began with a sound check, which also served as an attention check used as a basis for exclusion (participants had to read the text carefully to respond correctly). Participants were told that they were an astronaut-scientist on a recently discovered planet and that their job was to catalogue and describe new kinds of plants, animals, and objects that had been discovered on this new planet. Upon entering the lab, the participants encountered another scientist already working there. In each of three trials, the scientist introduced one of the novel categories and either intentionally or accidentally shared information about the features of one, two, three, or four exemplars. After the presentation, participants were asked a version of an *implied prevalence* question (Gelman et al., 2002; Cimpian et al., 2010; Tessler & Goodman, 2019b): "Imagine that you have another {*fep, dax, blicket*}, what are the chances it {*has white wings, has a black center, squeaks*}?" Participants responded using a slider bar with endpoints labeled 0% and 100%.

Participants were randomly assigned to one of ten conditions that differed in the manner in which the scientist communicated this information about the novel categories (*accidental examples* vs. *pedagogical examples*) crossed with the number of exemplars participants observed (1-to-4); in addi-

[1] Of course, the pedagogical context must signal an event wherein the teacher is aiming to inform the learner about the category and not, say, about a special member of the category.

[2] It should be noted that the account of Csibra and Shamsudheen (2015) proposes no direct or indirect link to be applied to adult cognition or even the cognition of young children who have acquired their first language. Thus, our argument should be understood as an application of the account of Csibra and Shamsudheen (2015) and not a direct theoretic consequence of it.

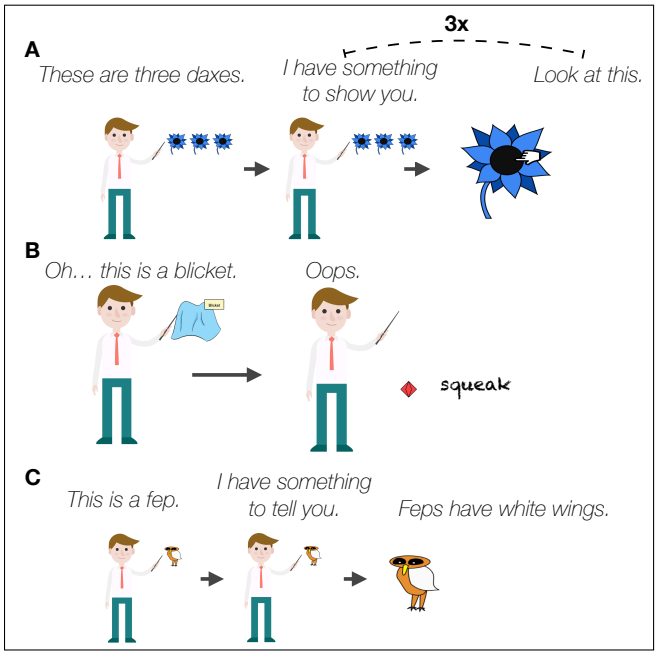

Figure 2: Overview of three conditions of the experiment, each showing a different item. A: 3x Pedagogical Example. The demonstration of the feature repeats 3 times. B: 1x Accidental Example. The speaker is learning about the object in the experiment (the object appears labeled, but hidden underneath a cloth). C: Generic + Pedagogical Example. Generic statement along with an intentionally demonstrated example.

tion, we include a *Generic Only* condition and a *Generic + Pedagogical Example* condition (Figure 2).

In the **Pedagogical Example** conditions, the scientist named the visually-displayed exemplar (e.g., "This is a fep.", "These are two feps.", etc.) and communicated about a feature in a pedagogical manner ("I have something to show you. Look at this!"). For the natural-kind categories (bird, flower), the image of the exemplar then enlarged while a white cursor-hand appeared to point to the feature of interest (white-wing, black-center, respectively; Fig. 2A); for the artifact, the object appeared to fall and make a squeaking sound.

In the **Accidental Example** conditions, the exemplar appeared underneath a blanket with a label attached (Fig. 2B). The scientist uttered: "Oh, this is a fep/blicket/dax" to indicate that he was learning about the object identity at that moment (presumably, via the label). The blanket then disappeared to reveal the feature. For the natural kind categories, the exemplar enlarged and the scientist remarked, "Oh, look at that!", expressing mild surprise (no cursor-hand pointed out the feature). For the artifact category, the scientist said "Oops" as the object fell and made a squeaking sound.

In the multiple exemplar conditions (2x, 3x, and 4x conditions), the exemplars were identical and the sequence of events repeated identically for each of the exemplars (e.g., speaker again saying "I have something to show you. Look at this." and demonstrating the feature, Fig. 2A). The scientist's utterances were presented both visually and auditorally in order to convey prosody information to reinforce the pedagogical vs. accidental manipulation (e.g., with surprise in the accidental condition). In neither accidental nor pedagogical conditions did the scientist explicitly label the feature.

The **Generic + Pedagogical Example** condition was identical to the Pedagogical Example condition, but with the speaker uttering a generic, saying "I have something to tell you. Feps have white wings / Daxes have black centers / Blickets squeak." (Fig. 2C). The **Generic Only** condition was an entirely text-based experiment, with the same cover story. Participants completed three trials of the same condition – manner of communication and number of exemplars were fixed across trials but each trial introduced a different category (order randomized). In other words, the manner of communication and number of exemplars were between-subject variables; category-type was a within-subject variable.

After the three main trials, participants completed a memory check trial. They were asked to select an exemplar for each of the three categories (e.g., "pick out the fep") from an array with three distractor items. Participants who failed to correctly identify all three category-types were excluded.

Results 16 participants failed to pass the attention/sound check, and 39 participants failed to correctly identify the exemplars during the memory check trials, which resulted in 347 participants for the main analyses. We observe a number of interesting qualitative features of the data, which exhibit substantial by-condition variability (Fig. 3). The first noteworthy feature is that a generic is worth more than a single observation, even one presented pedagogically (Generic vs. 1x Pedagogical). It is additionally remarkable that we find that the Generic + (1x) Pedagogical Example condition yields stronger generalizations than the Generic only condition. The strength of the generalization implied by a generic is enhanced with a concrete pedagogical example.

Second, we observe an interesting bi-modality in the responses for the low number of observations (1x or 2x) conditions. Many participants placed a fair bet (50%) that the next instance of the category will have the property, while others think that it is more likely than not (ratings of 60%-90%). This bi-modality persists with two accidental observations, but disappears after two pedagogical observations.

Bayesian analysis The question of how many observations one generic is worth is a natural question from a Bayesian hypothesis testing framework, where one can quantify the amount of evidence in support of a null hypothesis that two distributions are in fact the same (i.e., evidence in support of no-difference between conditions).[3] We do this by comput-

[3] In order to faithfully model the distribution of responses in each of the experimental conditions, we first performed a Bayesian analysis to determine the best function that characterizes our response variable, since they are clearly not normally distributed. We selected from a family of mixture of Beta distributions and determined that the data was much more likely to come from a mixture of Beta distributions than a single distribution (Bayes Factor BF $\approx 10^{20}$), though

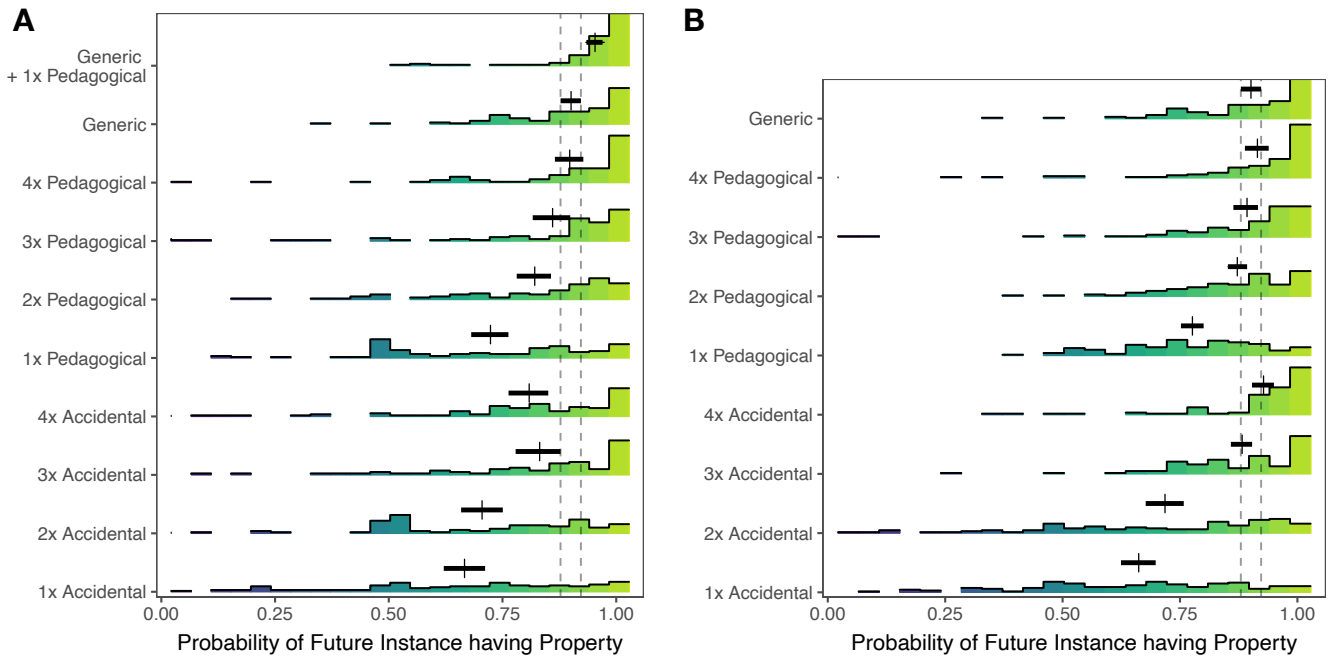

Figure 3: Experiment results. A: Experiment 1, the property was not labeled in the observation conditions. B: Experiment 2, the property was labeled in the observation conditions. Histograms of means and 95% bootstrapped confidence intervals appear above the empirical histograms. Dotted lines represented the confidence interval for the generic only condition. Data from the generic only condition of Experiment 1 is reproduced in (B) to ease visual comparison.

ing the marginal likelihood of the combined data set of the Generic condition and each of the other experimental conditions under the assumption that they are generated from the same distribution. We compare this likelihood to that calculated by assuming the two conditions were generated by independent distributions. The comparison of these marginal likelihoods gives us the Bayes Factor quantifying the evidence in support of the hypothesis that two conditions were generated from the same underlying distribution (i.e., the generic is worth n pedagogical or accidental observations).

We model the data for each condition independently as a mixture of two Beta distributions. We parameterize the Beta components by their mean μ and concentration ξ parameterization, and the Beta components i are combined via a mixture parameter ϕ. We put the following priors over the parameters: $\mu_i \sim \text{Uniform}(0,1)$, $\xi_i \sim \text{Exponential}(1)$, $\phi \sim \text{Uniform}(0.1)$. To compute the marginal likelihoods of the data for each model, we used an Annealed Importance Sampling algorithm (Neal, 2001) implemented in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014).

We find strong evidence against the hypothesis that a generic is worth a single positive observation, even one presented pedagogically (Table 1). The strongest evidence is that 4 pedagogical examples is worth the same as a generic, but already at 3 pedagogical examples we do see strong evidence for the equivalence. Interestingly, at no point do the accidental examples convincingly suggest they are equal to a generic.

Exploratory item analysis We see some evidence that generalization strength depends upon the category-type of the item, primarily in the 1x-3x accidental observations conditions. (Fig. 4A). Participants drew the strongest generalizations about the artifact and the weakest about the bird.

In addition to the artifact vs. natural kind distinction, our artifact examples were paired with an auditory property (*squeaking*), the demarcation of which is relatively explicit in both the pedagogical and accidental conditions. By contrast, our bird and flower items had visual properties (white features, black centers) which are not segregated from any other visual feature of the item. That is, the artifact's property is conveyed in a way that makes it clear what property to pay attention to, even though the speaker did not explicitly demarcate the property with words. This item difference illustrates one subtlety in comparing learning from observations to learning from generics: Generic statements explicitly articulate the property that a learner should attend to as well as potentially carry some core generic meaning that conveys generalization (i.e., *gen* in the semantic sense).

Experiment 2

One way in which generic language can foster generalization is by individuating a feature to be generalized. In Experiment 1, the feature being demonstrated was never individuated by labeling (i.e., the demonstrator just said "Look at this" and

the data was inconclusive as to whether or not it was a mixture of two distributions or of three (BF = 0.64).

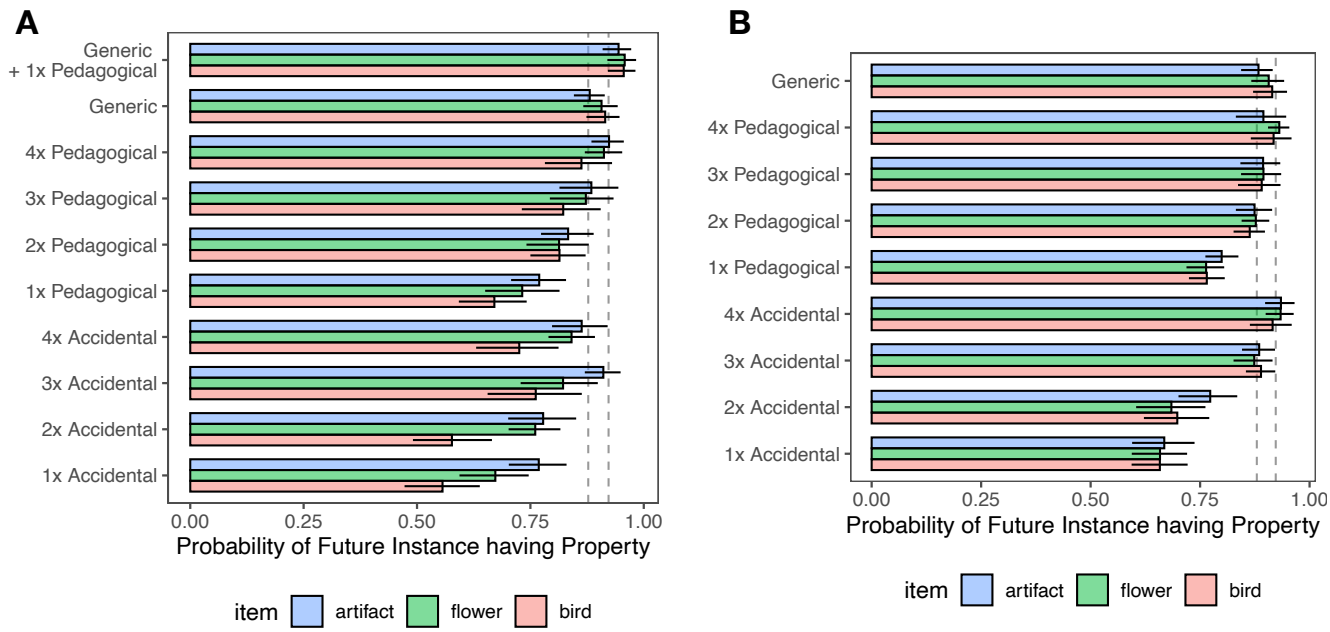

Figure 4: Experiment mean ratings broken down by item. A: In Experiment 1, participants drew stronger generalizations about the artifact than the other two items, primarily in the 1x-3x Accidental conditions. B: No comparable effect is observed in Experiment 2. Error-bars denote bootstrapped 95% confidence intervals.

a feature was either pointed to or in the accidental case, not indicated at all), leaving ambiguity about the exact feature being pedagogically highlighted or accidentally observed. To control for the possibility that individuating a feature is what enabled stronger generalizations from a generic than from a single observation in Experiment 1, we ran a follow-up experiment in which the learning events involving observations also included the labeling of the property.

Comparison	BF (Expt. 1)	BF (Expt. 2)
1 Accidental	4×10^{-12}	4.1×10^{-18}
2 Accidental	1.1×10^{-8}	6.1×10^{-9}
3 Accidental	33	8.1×10^{3}
4 Accidental	2.1	1.9×10^{3}
1 Pedagogical	2.6×10^{-7}	2.3×10^{-9}
2 Pedagogical	2.8	7.3×10^{2}
3 Pedagogical	1.2×10^{2}	1.5×10^{3}
4 Pedagogical	1.8×10^{3}	3×10^{3}
Generic + 1 Pedagogical	98	–

Table 1: Bayes Factors (BF) in support of the hypothesis that the strength of generalization implied by a generic is equal to that of the experimental condition.

Participants and procedure We collected data from 378 participants recruited from Amazon Mechanical Turk. For this experiment, we modified the Example conditions from Experiment 1, so participants were either assigned to the Pedagogical Example or Accidental Example condition and observed 1, 2, 3, or 4 exemplars (i.e., 8 conditions total). The Example conditions from Experiment 1 were modified such that the scientist provided both the label for the category and the name of the feature. In the Pedagogical Example, after naming the category, he said, "I have something to show you" and then named the feature: after the screen zoomed in on the bird or flower, he said "White Wings" / "A black center"; after the artifact dropped, he said "squeakin". In the Accidental Example, the scientist said, "Oh, Look at that! White wings/A black center" or "Oops! Listen to that! Squeaking!"

Results 39 participants were excluded for failing the memory check trials, resulting in 339 participants for the main analysis. Fig. 3B shows the responses for each condition, with the data from the Generics Only condition of Experiment 1 copied over for easier comparison. Foremost, we see that even when the property is explicitly demarcated by labeling, the strength of generalization from a single example – even pedagogically demonstrated – is not equal to that of a generic. Consistent with the findings in Experiment 1, we see that the change in generalization strength with increasing examples differs across the Pedagogical vs. Accidental Example conditions: The bi-modality in the distribution of responses disappears after 3 observations for the Accidental condition and only after 2 observations for the Pedagogical condition. We also see that with the property labeled, a generic is worth about 2 pedagogical examples or 3 accidental examples. We confirm these observations using the same Bayesian analysis as in Experiment 1 (Table 1). Finally, consistent with the idea that the artifact in Expt. 1 led to stronger generalizations

because the feature was clearly demarcated, we do not see appreciable differences between the items when the feature is labeled for all items (Fig. 4B).

Discussion

Successfully navigating the environment requires anticipating what is to come, and abstract generalizations allow us to reason flexibly about instances of categories and events that we have not yet experienced. These generalizations can be constructed both by directly observing instances in the world and by being told the generalization in the form of a generic sentence. But what is the relationship between learning from examples and learning from generics? Here we ask a simple question: How many observations is one generic worth? We find that, contra extant theoretical proposals, the strength of the generalization implied by a generic is equivalent to at least two pedagogically-sampled examples.

In our second experiment, we found evidence that describing the feature explicitly (e.g., "white wings") led to stronger generalizations than not describing the features with language. This points to an interesting dissection of the content of the generalization implied by generics. Part of the content of the generalization comes from simply articulating the feature. Interestingly, it is difficult to articulate a kind label and a feature label without conveying a generic. Generics are one of the most primitive syntactic and semantic constructions: nearly anytime you put a category and a property label together, you can get a generic meaning (e.g., "A dog barks").

Our results suggest that the model of Tessler and Goodman (2019a), which has been independently validated to explain human judgments about a wide range of generic sentences, somehow makes the wrong prediction with respect to the number of examples a generic is worth. The model's literal meaning for a generic implies that generics update beliefs in an analogous way to a single, pedagogical example (Tessler & Goodman, 2019b), which we found here to not be the case. This mathematical relationship between generics and observations, however, is derived by assuming the truth-conditional threshold for the generic follows a uniform prior distribution (i.e., all values of the generic threshold are equally likely). A non-uniform prior on thresholds skewed towards higher values would translate to more observations than just one. In a single interaction, pragmatic reasoning can be used to infer that higher thresholds are more likely, because if the speaker was using a lower threshold, their utterance would not have been very informative (Lassiter & Goodman, 2017; Tessler & Goodman, 2019b). The posterior distribution over thresholds after hearing a generic would be non-uniform and skewed towards higher values; this posterior could then become the prior for the next generic heard, which could be cashed in for more observations than just one. We leave the proof of this relationship for future work.

In our experiment, we used novel categories that would plausibly be construed as subordinate-level categories (i.e., a fep is a type of bird) to isolate the contribution of the number of examples without concern as to the variability of the examples. The generics–examples exchange-rate will, in general, depend upon the level of abstraction of the category. Acquiring a generalization about a superordinate category (e.g., "Mammals are warm-blooded") from examples will be more difficult than the subordinate categories we used. To draw a strong generalization about mammals, a learner would benefit not only from more examples but from more diverse examples (e.g., bears, cats, whales, ...). The generics–examples exchange-rate is thus not just one-dimensional (number of examples); it should also take into account the heterogeneity and representativeness of the examples.

Our experimental method is similar to other studies investigating the interpretation of generics vis a vis examples or concrete statistics. Cimpian et al. (2010) compared the strength of generalization implied by a generic to the statistics of the feature (e.g., "30% of lorches have purple feathers") that led participants to endorse the generic (i.e., judge "Lorches have purple feathers" as true), finding that generics were interpreted more strongly than what one would expect given the statistical information that yielded generic endorsement. Kushnir and Gelman (2016) examined the strength of generalization after hearing generic language and then observing instances with/without the property (e.g., hearing "Blickets squeak" and observing 2/10 blickets squeak). Both of these paradigms indirectly measure the generics–examples exchange-rate. In Cimpian et al. (2010), the equivalence is derived via truth judgments of generics (i.e., at what point do people endorse generics?). In Kushnir and Gelman (2016), instances of the category that lack the property can be explained away by the speaker's level of trustworthiness, which in turn influences the meaning of the generic heard. By contrast, in our paradigm, we map the strength of generalization implied by observations and by generics onto a common scale: predictions about a future instance.

A limitation of our paradigm that we may not evoke uninhibited, automatic communicative reasoning. Rather, we embed the paradigm in a story book that depicts certain communicative acts (Clark, 2016). For example, the Accidental Example condition is not really an accident: We experimenters designed the task in order to depict an accident. Despite this, we find that participants interpret the evidence presented in the Accidental Example conditions differently than they do same evidence presented in the Pedagogical Example conditions, lending some credence to the manipulation. Note that the manipulations were all between-subjects, so that any representation of the conditions as different is not via explicit reasoning about the different conditions *per se*.

Language and observations are the informational backbone upon which we build our knowledge of the world. The exchange rate of about two or three pedagogical examples for a generic suggests that the language of generalizations can save an instructor scrambling to find a good demonstration and that a few good examples are worth about as much as anybody can describe in words.

Acknowledgments

The authors would like to thank Karen Gu for her integral contributions to programming the experiment and data collection. This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship Grant No. 1911790 awarded to M.H.T., a National Science Foundation Graduate Research Fellowship Grant awarded to S.B., and Army Research Office MURI Grant No. W911NF-19-1-0057.

References

Butler, L. P., & Markman, E. M. (2012). Preschoolers use intentional and pedagogical cues to guide inductive inferences and exploration. *Child Development*, *83*(4), 1416–1428.

Carlson, G. N. (1977). Reference to kinds in English. *PhD thesis, University of Massachusetts*.

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive science*, *34*(8), 1–30. doi: 10.1111/j.1551-6709.2010.01126.x.Generic

Clark, H. H. (2016). Depicting as a method of communication. *Psychological Review*, *123*(3), 324.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153.

Csibra, G., & Shamsudheen, R. (2015). Nonverbal generics: Human infants interpret objects as symbols of object kinds. *Annual review of psychology*, *66*, 689–710.

Gelman, S. A., Star, J. R., & Flukes, J. E. (2002). Children's Use of Generics in Inductive Inferences. *Journal of Cognition and Development*, *3*(2), 179–199.

Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2759–2764).

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* `http://dippl.org`. (Accessed: 2020-5-20)

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological Review*, *111*(1), 3.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental psychology*, *37*(5), 620.

Gweon, H., & Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, *332*(6037), 1524–1524.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066–9071.

Kushnir, T., & Gelman, S. (2016). Translating testimonial claims into evidence for category-based induction. In *Cogsci.*

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, *194*(10), 3801–3836.

Leslie, S.-J. (2007). Generics and the structure of the mind. *Philosophical perspectives*, *21*, 375–403.

Meyer, M., & Baldwin, D. A. (2013). Pointing as a sociopragmatic cue to particular vs. generic reference. *Language Learning and Development*, *9*(3), 245–265.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and computing*, *11*(2), 125–139.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309–318.

Tessler, M. H., & Goodman, N. D. (2019a). The language of generalization. *Psychological review*, *126*(3), 395.

Tessler, M. H., & Goodman, N. D. (2019b). *Learning from generic language.* PsyArXiv. Retrieved from `psyarxiv.com/hnm8p`