Cognition 244 (2024) 105707

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Testimony and observation of statistical evidence interact in adults' and children's category-based induction

Zoe Finiasz [a,*], Susan A. Gelman [b], Tamar Kushnir [a]

[a] Department of Psychology & Neuroscience, Duke University, 417 Chapel Drive, Box 90086, Durham, NC 27708, United States of America
[b] Department of Psychology, University of Michigan, 530 Church Street, Ann Arbor, MI 48109, United States of America

ARTICLE INFO

Keywords:
Generics
Causal learning
Cognitive development
Testimony
Statistical learning

ABSTRACT

Hearing generic or other kind-relevant claims can influence the use of information from direct observations in category learning. In the current study, we ask how both adults and children integrate their observations with testimony when learning about the causal property of a novel category. Participants were randomly assigned to hear one of four types of testimony: generic, quantified "all", specific, or only labels. In Study 1, adults (N = 1249) then observed that some proportion of objects (10%–100%) possessed a causal property. In Study 2, children (N = 123, M_{age} = 5.06 years, SD = 0.61 years, range 4.01–5.99 years) observed a sample where 30% of the objects had the causal property. Generic and quantified "all" claims led both adults and children to generalize the causal property beyond what was observed. Adults and children diverged, however, in their overall trust in testimony that could be *verified* by observations: adults were more skeptical of inaccurate quantified claims, whereas children were more accepting. Additional memory probes suggest that children's trust in unverified claims may have been due to misremembering what they saw in favor of what they heard. The current findings demonstrate that both child and adult learners integrate information from both sources, offering insights into the mechanisms by which language frames first-hand experience.

It's nearly impossible to learn something new without in some way generalizing what you learned. This core feature of cognition – that our experiences provide information beyond just individual instances – supports the inferences, predictions, and explanations that help us make sense of the world around us. Consider the following example. Professor X just visited an exotic island, and informs you that Ylang-Ylang flowers, which can only be found on this island, are fragrant. A reasonable expectation, based on this testimony, is that when you visit the island yourself, most Ylang-Ylang flowers you would encounter would be fragrant. But what happens when you discover on your island visit that most of the Ylang-Ylang flowers have no smell? You may be inclined to wonder if Professor X was mistaken, or – if you trust that he wasn't – perhaps your own observations are not representative of Ylang-Ylang flowers in general. You may wonder, 'Did I smell a bad batch of flowers, or was the professor wrong?'

This example illustrates two different sources of evidence we can use to learn about kinds: observation and testimony. In two studies, we aim to show the ways in which different kinds of testimony lead to different expectations, shaping how both adults and children use subsequent probabilistic observations in category learning. We will begin by discussing how learners use both testimony and observation to form beliefs about categories. Following this, we will discuss the potential ways in which these two sources interact, and implications for category learning when testimony is offered prior to observing probabilistic evidence first-hand.

One way to gather first-hand evidence about categories is through observation by sampling from the category and observing the prevalence of that property in the sample (i.e., the statistical likelihood of a member of the sample having said property), then treating the sample as representative of the population.[1] Much empirical research – guided by formal models based on Bayesian inference – suggests that generalizations from samples to populations are principled and rational (Gopnik &

* Corresponding author.
E-mail addresses: zoe.finiasz@duke.edu (Z. Finiasz), gelman@umich.edu (S.A. Gelman), tamar.kushnir@duke.edu (T. Kushnir).

[1] We leave aside for now, specifics on the way the sample was generated. In our Ylang-Ylang example, it could have been chosen randomly (he happened to pick some flowers), chosen intentionally (he liked those flowers specifically, so he picked them), or chosen pedagogically (he wanted to share information about their fragrant properties, so he picked a representative set to show me). For details on how rational models account for the sample-generating process in inductive generalization, see Shafto et al. (2012); Shafto et al. (2013).

https://doi.org/10.1016/j.cognition.2023.105707
Received 1 June 2023; Received in revised form 16 October 2023; Accepted 14 December 2023
Available online 3 January 2024

Wellman, 2012; Oaksford & Chater, 2007; Schulz, 2012; Tenenbaum & Griffiths, 2001); that is, they are based on our current conceptual knowledge and are appropriately responsive to new evidence. The rational modeling approach has empirical support from concept learning studies in infants (Denison, Trikutam, & Xu, 2014; Sobel & Kirkham, 2007; Teglas, Girotto, Gonzalez, & Bonatti, 2007; Xu, 2019), children (Gopnik & Wellman, 2012; Kimura & Gopnik, 2019), and adults (Griffiths, Sobel, Tenenbaum, & Gopnik, 2011; Griffiths & Tenenbaum, 2009; Kemp & Tenenbaum, 2009).

Another source of evidence is testimony. Much of our conceptual knowledge is acquired by listening to what people tell us, because many of the deep, non-obvious, and essential properties of categories are not directly observable (Gelman, 2003, 2023; Harris, 2012; Harris & Koenig, 2006). We therefore rely on others to impart generalities by making kind-based claims (Cimpian & Markman, 2009; Gelman et al., 1998; Koenig et al., 2015). To this end, languages make use of two types of devices to refer to kinds as opposed to individuals: generic noun phrases (e.g., "*Birds* have hollow bones") and quantifiers (e.g., "*All birds* have hollow bones"; "*Most birds* have hollow bones").

Research has shown that learners can use both generics and quantifiers as the basis for probability estimates (Brandone, Gelman, & Hedglen, 2015; Cimpian, Brandone, & Gelman, 2010; Cimpian, Gelman, & Brandone, 2010; Gelman, Star, & Flukes, 2002; Rhodes, Gelman, & Brickman, 2010; Tessler, Bridgers, & Tenenbaum, 2020). but that they differ in important respects. Quantifiers have precise semantic implications that can readily be translated into probabilities: "all" expresses 100%, "most" expresses >50%, and "some" expresses a non-zero quantity (often <50%, due to scalar implicatures; Noveck, 2001). Generics, by contrast, systematically and precisely differ from probabilistic representations (Cimpian, Brandone, & Gelman, 2010, Cimpian, Gelman, & Brandone, 2010; Gelman & Bloom, 2007; though see Tessler et al., 2020). For one, generics do not reflect a consistent frequency but rather express features that are conceptually central to a kind (Butler & Markman, 2014; Cimpian & Markman, 2009). Thus, the frequency of a generic proposition may range from all members of a category ("Giraffes are ungulates") to roughly half the category ("Lions have manes") to <1% of members ("Mosquitoes carry the West Nile Virus"). Furthermore, generics have several semantic properties that distinguish them from quantifiers: they permit exceptions (e.g., "Birds fly" does not imply that all birds fly; Hollander, Gelman, & Star, 2002) and imply broader prevalence than is required for their use (Brandone et al., 2015; Cella, Marchak, Bianchi, & Gelman, 2022; Cimpian, Brandone, & Gelman, 2010; Cimpian, Gelman, & Brandone, 2010). Thus, unlike claims made with quantifiers, generic claims make members of a category seem more alike. This can be beneficial when the goal is to learn broad patterns in a group but may also lead learners to overlook individual differences and to treat instances as standing in for the group to which they belong.

To summarize, observations and testimony each have benefits and drawbacks for category learning. Observations of statistical evidence can be used to make rational inductive inferences about observable properties of category members, but learning is restricted to properties that are observable, and requires assuming that samples are representative of populations. Further, although observations have the benefit of being direct evidence, they are limited in the case of small sample sizes (Rhodes, Brickman, & Gelman, 2008). Testimony – generic language in particular – is useful for learning about unobservable properties of category members and for highlighting feature centrality but cannot be mapped on to a precise quantity (Tessler et al., 2020). Furthermore, some of the semantic properties of generic language suggest that hearing generic testimony in combination with observations may lead to overly strong assumptions about the representativeness of small samples (Rhodes, Leslie, & Tworek, 2012). Testimony also has the benefit of being able to describe entire populations and thus can provide information about a wider scope of category members but may be more difficult to verify.

Here we investigate the possibility that, when testimony is available prior to first-hand observations, the two sources mutually inform each other. People are often in the position of hearing about categories from those who claim to be knowledgeable (as in our traveling professor example) but then later on may have the opportunity to *verify* this information for themselves with their own observations. What happens in such cases depends on the testimony they hear, and how this informs their expectations of what they will see. In fact, both kind-based claims *and* instance-specific claims may have this effect: when one hears, "*This* Ylang-Ylang flower is fragrant," one might expect to observe that a single Ylang-Ylang flower is distinctive, and that the general tendency of such flowers is not to be fragrant (Cimpian, Brandone, & Gelman, 2010; Cimpian, Gelman, & Brandone, 2010). On the other hand, prior work has shown that hearing generic claims, such as "Ylang-ylang flowers are fragrant," leads to the expectation that the property is either ubiquitous, or dangerous, or a central feature of the category (Cimpian, Brandone, & Gelman, 2010; Cimpian, Gelman, & Brandone, 2010; Gelman, Ware, & Kleinberg, 2010; Rhodes et al., 2012). Because learners expect that generic claims refer to high-prevalence properties, hearing such claims may have different effects when followed by observations that match these expectations (observing the property with high frequency) than when followed by observations that mismatch expectations (observing the property with low frequency).

A study by Chambers and colleagues (2008) shows that the same sample of deterministic observations leads to different patterns of generalization depending on whether it follows generic or non-generic claims. Here, adults and children were presented with either generic or non-generic testimony about a novel category (i.e., "Pagons are friendly" or "This pagon is friendly") and followed this with observations of deterministic evidence of the property in a sample (100%, or 5 out of 5 pagons, were friendly). Learners of all ages were more likely to generalize a property to a single new category member after hearing generic testimony than specific or non-generic testimony (also see Hermansen, Ronfard, Harris, Pons, & Zambrana, 2021; Hoicka, Saul, Prouten, Whitehead, & Sterken, 2021; Stock, Graham, & Chambers, 2009).

This example leaves open our original question about the professor and his flowers. When testimony sets up expectations that are *not* met by first-hand observations, how do learners resolve this conflict? Prior work shows that even young children check testimony against their own observations when inferring speaker reliability (Koenig & Harris, 2005a, 2005b; Mills, 2013), knowledge (Sobel & Kushnir, 2013), expertise (Keil, 2010; Kushnir, Vredenburgh, & Schneider, 2013), and general trustworthiness (Harris & Corriveau, 2011). Moreover, mismatches between testimony and children's own observations in a categorization paradigm led participants to show decreasing reliance on the testimonial claims over time (Ronfard, Lane, Wang, & Harris, 2017). Thus, when expectations formed by hearing testimony are not matched by one's own observations, learners have shown a tendency to dismiss or mistrust claims due to skepticism of the speaker's credibility.

Yet kind-based claims, especially generic claims, present a unique case. Unless they strongly contradict our own knowledge or observations (to continue our example, this would be the equivalent of being certain that *no* Ylang-Ylang flowers have a fragrance), they are not easy to verify (because the kind is a non-visible, abstract entity; Gelman, 2023), and thus potentially not easy to dismiss. In support of this, a study by Koenig et al. (2015) showed that child and adult learners explicitly rate speakers who make generic claims about kinds as more knowledgeable than those who make specific claims about individual members of a kind, even when those claims are not easily verifiable. Since observing a small sample of instances – especially probabilistic evidence – can yield uncertainty about representativeness, this could be enough to lead learners to weigh claims from a speaker that they consider to be knowledgeable more heavily than their own observations.

The review above suggests a pathway through which observations and testimony may interact to influence category learning. Testimony sets up expectations that either can be verified by a sample of

observations or can conflict (partially or wholly) with the sample. When there is conflict, we predict two consequences. First, claims that are hard to verify (such as generic claims) might weigh more heavily in category-based induction than first-hand observations of statistical evidence. But also, observations can help a learner evaluate a speaker's claims to know, and, to the extent that learners judge that speaker to be more (or less) knowledgeable, the claims may be given more (or less) weight in inductive generalizations.

Our method is a category learning task in which objects of a novel kind (e.g., blickets) had a causal property (making a machine light up and play music) with some probability, ranging from 10% to 100%. In Study 1, we examine the degree to which adults generalize the causal property to new objects and attribute knowledge to the speaker. We then explore category-based inductive inferences both when evidence supports testimony and when it conflicts, and whether inferences are mediated by ratings of speaker knowledge. We follow this with Study 2, a simplified version of the task with young children that emphasizes cases of conflict between kind-based claims and low-frequency statistical evidence. The work reviewed above suggests parallels between children and adult learners both in learning from statistical evidence and in use of kind-referring testimony. Thus, we might expect children to integrate both sources of evidence in similar ways to adults. On the other hand, there are reasons to expect differences due to the unique power of testimony to shape children's beliefs and causal inferences, and even to bias their memories for observed events. More details on these potential differences are discussed in Study 2. Stimuli and data for both studies can be found at https://osf.io/d73jx/.

1. Study 1

We presented a novel object category (e.g., blickets) and demonstrated a property (making a machine light up and play music) of a sample of category members, with some probability. Participants were randomly assigned to conditions that systematically varied the frequency of the property from 10% (1 out of 10 objects in the sample had the property) to 100% (all of the 10 objects in the sample had the property). One group of learners, randomly assigned, were taught the name of the category only, and observed the evidence in the absence of any testimony about the property (*Label-only* condition). In this case, we expected that estimates of the prevalence of the property in the category would increase as a function of an increase in frequency of the property in the sample.

We also tested whether this relation between observed frequency and prevalence estimates changed when introducing different types of testimony in three experimental groups, randomly assigned: Generic, Quantified (*All*), or Specific (*This*). One set of participants in each observed frequency condition (10 total groups) heard an informant make a Generic claim about the property (e.g., "Blickets make the machine go"). Another set of participants (again, 10 total groups, one in each observed frequency condition) heard a Quantified claim that the property was present in "all" category members (e.g., "All blickets make the machine go"). We compared these two sets of participants who heard kind-based claims about the property (i.e., generic or 'all') with one set (10 groups) who heard the informant make a Specific claim about the property of only one of the objects in the sample (e.g., "This blicket makes the machine go"). As explained above, we contrasted all three testimony conditions with a control group that only heard an informant label the objects and was given no information about the properties (label-only condition).

Our measure of interest was participants' property prevalence estimates – do kind-based testimonial claims influence prevalence estimates across observed frequencies? If, as has been found in prior work, participants assume that generic and quantified "all" claims indicate high prevalence, then perhaps those who hear either type of claim would over-estimate the prevalence of the property above baseline, and above hearing specific claims about the property of one category member. However, given semantic distinctions between generic and quantified claims, the relation between observed frequency and prevalence estimates might also differ between the two conditions. Because generic claims are judged to be true for low-frequency properties as well as high-frequency ones, this could lead to different prevalence estimates for "all" than for generic wording, across the full range of observed frequencies. We contrast this with the other two conditions, the Specific ("This") and Label-Only conditions, where prevalence estimates are expected to match observed frequencies more closely.

Our second question was how participants would use their observations to verify testimonial claims, and, based on this, make judgments about the knowledge of the informant. To explore this, we asked for ratings of the informant's knowledge of the category and of the causal property, and we also asked participants if they endorsed additional claims made by the informant of a novel category label, and a novel claim about the causal property. We were particularly interested to see how participants would react to kind-based claims that mismatched observed frequencies (i.e., low observed frequency). Under conditions of mismatch, we expected participants' attributions of speakers' knowledge, and endorsements of novel claims, to be low. We reasoned that low knowledge attributions/endorsements would be most likely when observations logically contradict speaker's testimony (in the "All" condition, where testimony can be contradicted by a single counterexample). Generic claims paired with low observed frequency could lead to one of two patterns: Either they might lead to low knowledge attributions/endorsements because the testimony implies high prevalence, or, given the flexibility of generic claims to varying interpretations, knowledge attributions following such claims may remain high. We contrast this with how participants rate the knowledge of an informant who makes specific, verifiable claims that are limited in scope to one category member, and how they rate the knowledge of an informant who offers no information beyond a category label.

Our third question was whether we would find a relation between knowledge attributions and prevalence estimates. This could occur if attributions of knowledge influence estimates, and participants assign greater weight to any testimony they perceive to be coming from a knowledgeable source. This could also occur if estimates influence knowledge ratings, and participants rate a source as more knowledgeable when their claims match the participant's own assessment of events. Either way, a separate relation between knowledge attributions and estimates that holds regardless of the observed frequency would provide further support for the idea that participants use each source of evidence (observations and testimony) as a way to verify and check the other.

1.1. Method

1.1.1. Participants

One thousand two hundred and forty-nine adults (654 female, age range 16–86 years, $M_{age} = 37$ years, $SD = 11$ years) participated in the study through Amazon Mechanical Turk in exchange for monetary compensation. Data were collected from U.S. participants only. Participants were majority white (80%) and non-Latino/a (92.9%). A majority had attended college (38% with a 4-year degree, 12% with a 2-year degree, 26% some college, 12% with a graduate or professional degree).

1.1.2. Procedure

A random number generator assigned each participant to one of 40 conditions, for a total of 30–33 participants per condition. Each condition combined one of 4 types of testimonial claims (Specific, Generic, Quantified ("All"), Label Only) and one of 10 observed frequencies of the novel causal property (ranging from 1/10 to 10/10, inclusive).

The procedure consisted of two trials, each with the same characters, observed frequency, and testimony, but with different objects (Trial 1: "blickets" and Trial 2: "midos") and a different-colored machine. On each trial, participants first watched a short video then answered a set of

questions. The video began with a girl who introduced the "special music machine," her friend Zorg, and an array of objects on a shelf that she labeled (i.e., "Here are some blickets" in Trial 1). Zorg then made one of the following testimonial claims, depending on condition (see Fig. 1):

Generic claim: *"Wow, look at those blickets! I know something about blickets. Blickets make the machine go."*

Quantified ("All") claim: *"Wow, look at those blickets! I know something about all blickets. All blickets make the machine go."*

Specific claim: *"Wow, look at those blickets! I know something about this blicket. This blicket makes the machine go."* [In this condition, the blicket Zorg refers to is always the one closest to him, and as he speaks, he "points" to it with a dashed line connecting his hand to the object.]

Label Only [baseline]: *"Wow, look at those blickets!"*

After hearing Zorg's claim, the video showed the 10 objects passing through the machine (they entered the top and exited the bottom), one at a time. When the machine "activated," it changed color and made a loud trumpeting sound. The first object always had the property of activating the machine, insuring that, in the Specific condition, Zorg's claim was accurate. After the first object, a random subset of the remaining objects activated the machine, depending on the observed frequency assigned.

Following the video, participants made two knowledge attribution ratings, which together served as their evaluation of the knowledge of the speaker: A Category Knowledge Attribution (*"How much do you think Zorg knows about blickets?"*) and a Machine Knowledge Attribution (*"How much do you think Zorg knows about the machine?"*). Answer choices for both were on a three-point scale: 0 – Nothing, 1 – A little bit, 2 - A lot. Participants also answered the focal Property Prevalence Estimate: *"Imagine there were more blickets here. What percentage of these blickets would make the machine go?"* Participants were allowed to enter numbers ranging from 0 to 100. The trial concluded with two opportunities for participants to endorse claims made by the informant, each about a different novel object. The first was a New Label claim; Zorg labeled a novel object ("Lem"), and participants were asked, *"Is this a lem?"* The second was a Same Property claim; Zorg claimed that a new object "makes the machine go" (with the same machine as in the "blickets" movie) and participants were asked, *"Does this make the machine go?"* Responses were coded 1 = yes, 0 = no. Following Trial 1, participants began Trial 2 by watching the second video and answering the same set of questions about "midos." The observations and testimony were the same for each trial.

1.2. Results

1.2.1. Property prevalence estimates

We examined the influence of Trial Order (First or Second, within participants), Observed Frequency (1/10 through 10/10, between participants), Testimony (Specific, Generic, Quantified, Label Only, between participants), and the interaction of Observed Frequency and Testimony on Prevalence Estimates, using a linear mixed effects model.[2] Parameter estimates for the model are shown in Table 1, with Testimony dummy coded so that the Label Only condition is the reference category. The main effects and interaction can be seen in the model of predicted values in Fig. 2.

There was no effect of Trial Order ($F(1,2460) = 0.06$, *ns*). There was a significant main effect of Observed Frequency ($F(1,2460) = 3311.3$, $p < .001$), such that higher observed frequencies led to higher prevalence estimates. There was also a significant main effect of Testimony ($F(3,2460) = 11.9$, $p < .001$) such that prevalence estimates were on average higher in the Generic and Quantified ("All") conditions than in the Label Only condition. Finally, there was a significant interaction between Testimony and Frequency ($F(3,2460) = 6.6$, $p < .001$). In the Generic conditions, the slope of prevalence estimates was significantly smaller than the slopes of all three other conditions (Label Only, $t(3) = -4.15$, $p < .001$; Specific, $t(3) = 3.46$, $p < .001$; Quantified $t(3) = 2.70$, $p < .001$). The slopes did not differ between the other three conditions.

Fig. 3 suggests that generic language may have the effect of "flattening" participants' estimates at the extremes – that is, participants were more likely to estimate above observed frequencies when they were low, and more likely to estimate below observed frequencies when they were high. To check whether this was the case, we used one-sample *t*-tests to compare the observed to the estimated frequencies in each of the generic language conditions (from 10% up to 100%; see Fig. 3 for a plot of all of the means and 95% CI). These t-tests suggest that the effect of generic language was strongest at both extremes: In the 1/10 condition the average *overestimate* was 17.6% above observed frequencies ($M = 27.6\%$, $SE = 5.4\%$, $t(32) = 3.26$, $p < .01$), and in the 10/10 condition the average *underestimate* was 17.8% below observed frequencies ($M = 82.2\%$, $SE = 5.2\%$, $t(29) = -3.43$, $p < .01$). The remaining conditions were mixed: Participants in the 2/10 and 3/10 conditions overestimated (2/10: $M = 29.1\%$, $SE = 4.2\%$, $t(30) = 2.16$, $p < .05$; and 3/10: $M = 35.4\%$, $SE = 2.3\%$, $t(29) = 2.36$, $p < .05$). Participants in the 4/10 condition significantly underestimated ($M = 33.3\%$, $SE = 2.2\%$, $t(31) = -3.06$, $p < .01$). Participants in the 5/10, 6/10, 7/10, and 9/10 conditions estimated around the observed frequencies (*p*s > 0.05). Participants in the 8/10 conditions underestimated ($M = 71.2\%$, $SE = 4.2\%$; $t(30) = -2.09$, $p < .05$). While results should be interpreted with caution, they suggest participants overestimated low observed frequencies, and may have done so more often than they underestimated high ones.

1.2.2. Label and property endorsements

We analyzed the percentage of endorsements for each of the questions using a logistic GEE with Trial (1 or 2, within participant) and Endorsement Type (New Label, Same Property, within participant), Testimony (Label Only, Specific, Generic, Quantified, between participants), Observed Frequency (1/10 through 10/10, between participants), and the various interactions (all 2- and 3-way except with Trial) as predictors. There a main effect of Endorsement Type (Wald $\chi^2(1) = 72.5$, $p < .001$), a main effect of Testimony (Wald $\chi^2(3) = 17.4$, $p = .001$), a main effect of Observed Frequency (Wald $\chi^2(1) = 10.7$, $p = .001$), an interaction of Testimony and Observed Frequency (Wald $\chi^2(1) = 8.5$, $p = .037$), and no other significant effects.

Fig. 4 shows the results across conditions in detail. Overall, New Label endorsements – in which participants were asked to endorse Zorg's claims about labels for new objects – were higher than Same Property endorsements – in which participants were asked to endorse Zorg's claims about a new object having the same causal property. New Label endorsements were similar across Testimony conditions (Wald $\chi^2(3) = 5.52$, *ns*), but Same Property endorsements were different depending on Testimony condition (Wald $\chi^2(3) = 29.81$, $p < .001$). Fig. 4 shows that the interaction of Observed Frequency and Testimony was mainly due to Same Property endorsements at or near chance (50%) in the Quantified ("All") condition at low observed frequencies.

1.2.3. Knowledge attributions

We then examined the influence of Trial Order, Observed Frequency, Testimony, and the interaction of Observed Frequency and Testimony on average Knowledge Attributions (combined across both Category Knowledge and Machine Knowledge questions), using a linear mixed-effects model. Parameter estimates for the model are shown in Table 2; the Label Only condition is the reference category. An illustration of the main effects and interaction can be seen in Fig. 5.

There was no significant effect of Trial Order ($F(1,2488) = 0.2$, *ns*). There was a significant main effect of Observed Frequency ($F(1,2488) = 125.9$, $p < .001$). There was also a significant main effect of Testimony ($F(1,2488) = 22.2$, $p < .001$): attributions of knowledge were highest in the Specific condition ($M = 1.32$, $SD = 0.52$), next highest in the Generic

[2] A Q-Q plot of the prevalence estimates showed that they were approximately normally distributed.

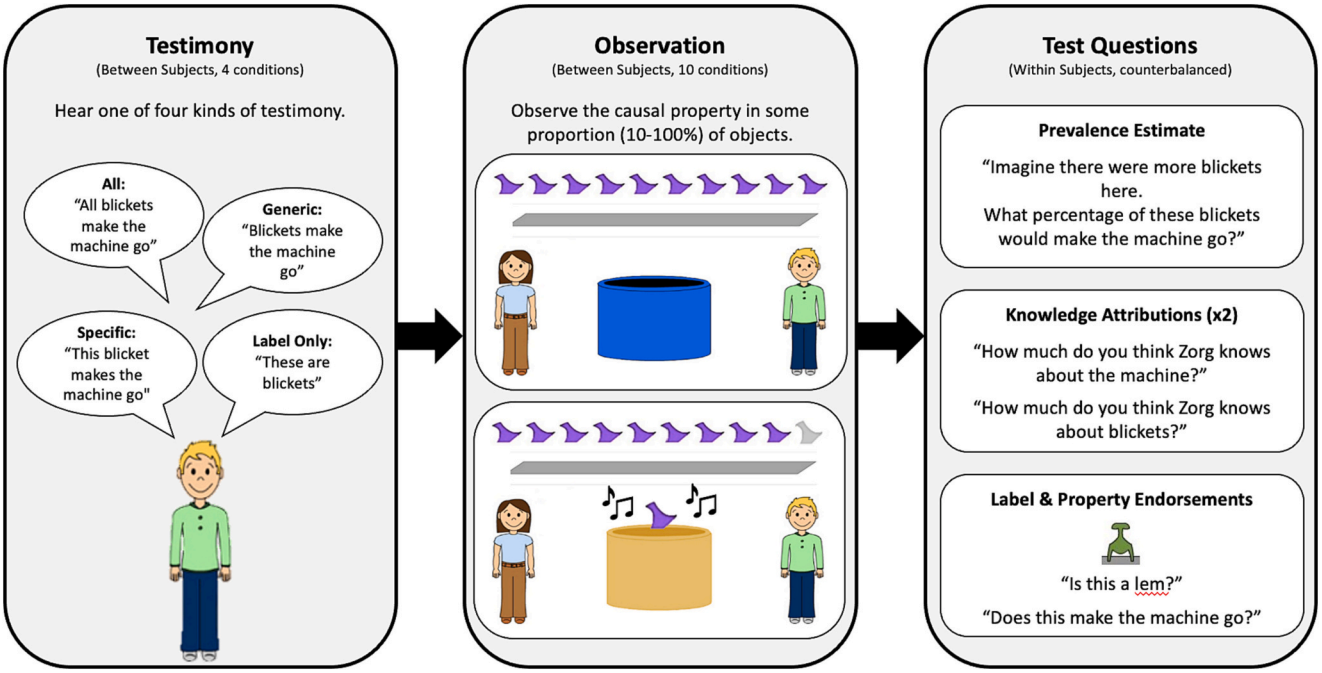

Fig. 1. Design of the procedure for Study 1. Participants were assigned to one of four testimony conditions and one of 10 frequency conditions for a total of 40 between-subjects conditions.

Table 1
Parameter Estimates of the mixed model predicting Prevalence Estimates by condition. Testimony is dummy coded so that the Label Only condition is the reference category.

Parameter	Estimate	SE	*t (df = 2460)*
Intercept of Label Only condition	3.70	1.80	2.05***
Slope of Label Only condition	0.89	0.03	31.25***
Intercept difference between Label Only and Specific Conditions	0.17	2.50	0.07
Intercept difference between Label Only and Generic Conditions	11.72	2.48	4.73***
Intercept difference between Label Only and Quantified ("all") Conditions	9.24	2.50	3.70***
Slope difference between Label Only and Specific Conditions	−0.03	0.04	−0.67
Slope difference between Label Only and Generic Conditions	−0.17	0.04	−4.15***
Slope difference between Label Only and Quantified ("all") Conditions	−0.06	0.04	−1.39

Adj. $R^2 = 0.62$.
*$p < .05$.
**$p < .01$.
***$p < .001$.

condition ($M = 1.22$, $SD = 0.51$), and lowest in the Quantified ($M = 1.07$, $SD = 0.53$) and Label Only ($M = 1.04$, $SD = 0.46$) conditions. Finally, there was a significant interaction between Observed Frequency and Testimony ($F(1,2488) = 9.9$, $p < .001$). The correlation between knowledge attributions and observed frequency were significantly greater in the Generic and Quantified conditions than in the Specific and Label Only conditions (see parameter estimates in bold in Table 2).

1.2.4. Knowledge attributions and prevalence estimates

Our final question concerned whether knowledge attributions play an independent role in predicting prevalence estimates above and beyond the influence of Observed Frequency and Testimony. We addressed this question by regressing the average Knowledge Attributions and their interactions with Observed Frequency and Testimony on the *residuals* from the model in Table 1. The main effect of Knowledge Attributions was significant ($F(1,2460) = 60.0$, $p < .001$) as were the interactions between Knowledge Attributions x Testimony ($F(3,2460) = 7.7$, $p < .001$), between Knowledge Attribution x Observed Frequency ($F(1,2460) = 10.3$, $p = .001$), and the three-way interaction ($F(3,2460) = 4.6$, $p = .003$).

In order to interpret the interactions, we looked at the correlations of prevalence estimates and the deviations from condition averages (residuals of the model in Table 1) as a function of Testimony. In the No Label and Specific Conditions, deviations of prevalence estimates from the condition average were not related to knowledge attributions (correlation between residuals and average Knowledge Attributions: No Label condition, $r^2(314) = -0.06$, *ns*; Specific Condition, $r^2(311) = 0.06$, *ns*). This is unsurprising, given that knowledge attribution ratings in these conditions were not substantially related to observed frequency (see Table 2/Fig. 5).

In the Generic and Quantified conditions, higher knowledge attributions related to higher-than-average prevalence estimates (correlation between residuals and Average Knowledge Attributions: Generic condition, $r^2(312) = 0.24$, $p < .001$; Quantified Condition, $r^2(307) = 0.28$, $p < .001$). Bearing in mind that most participants in the Generic and Quantified conditions gave higher knowledge ratings at high observed frequencies (see Fig. 5), this result suggests that the highest knowledge ratings (i.e., participants who said Zorg knew "a lot" rather than just "a little") corresponded to higher-than-average prevalence estimates.

1.3. Discussion

In Study 1, adult participants were provided with two types of evidence about a novel category and its property. They heard a testimonial claim, and they observed the property with some frequency in a sample of category members. Provided with this evidence, participants were asked to estimate the prevalence of the property in the category as a whole, and to attribute knowledge to the informant who made the testimonial claim.

To begin with, it should be noted that participants in our study responded in a systematic way to the statistical information in the sample they observed, which is consistent with prior work on causal

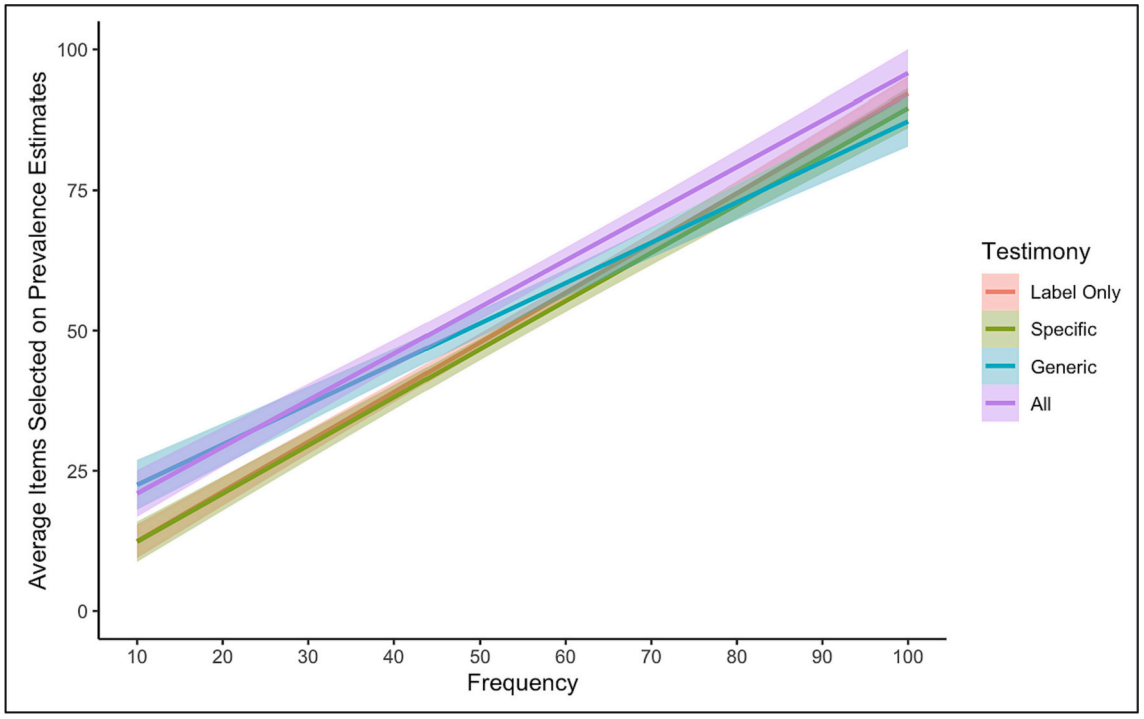

Fig. 2. An illustration of the main effects of Frequency, Testimony, and the interaction between the two on participants' estimates of property prevalence.

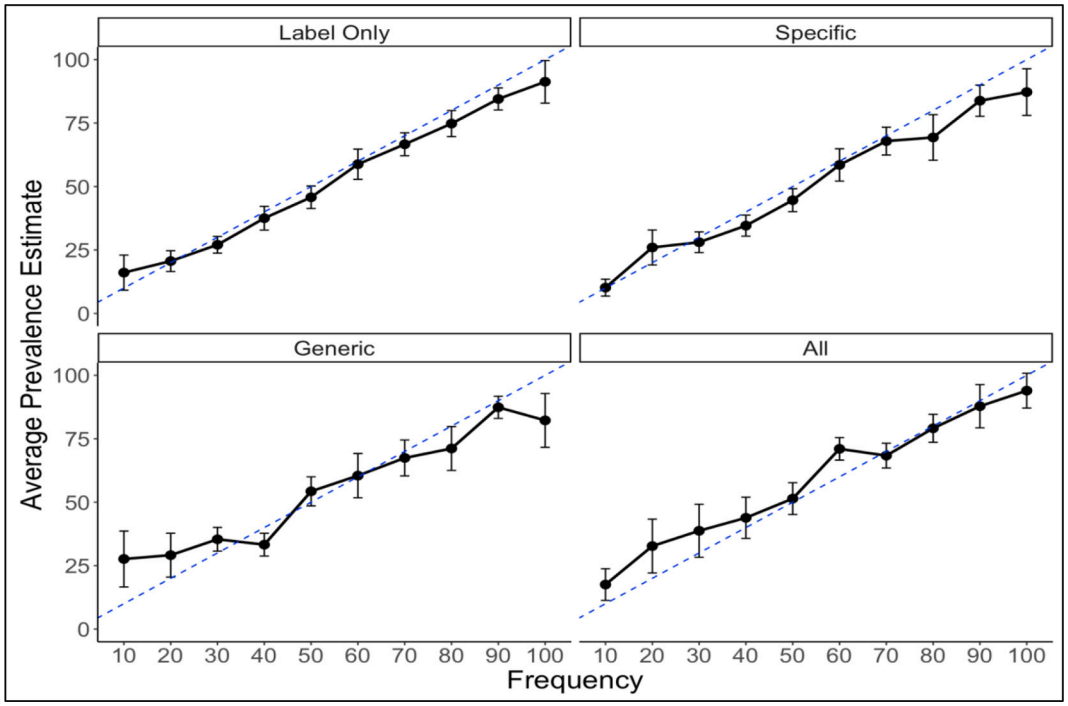

Fig. 3. Average prevalence estimate, and 95% confidence intervals of the mean by Testimony and Frequency conditions. The dashed line represents the actual observed frequency.

inference from statistical evidence. Low observed frequencies led to low prevalence estimates, and high observed frequencies led to high prevalence estimates. However, testimonial claims about the category influenced this basic pattern.

Following Generic claims ("*Blickets* make the machine go"), the linear relation between observed frequency and prevalence estimates was flatter than in the other three conditions. This effect can be seen most clearly at the extreme ends of the range. Following low observed frequencies (10–30%), participants' estimates were higher than the observed frequencies (Fig. 3), and relatively higher than when a speaker made no property claims (Label Only condition) or verifiable claims about a single category member (Specific condition). A comparable effect appears, but in the opposite direction, for high frequency cases (70%–100%). Moreover, attributions of knowledge to the generic

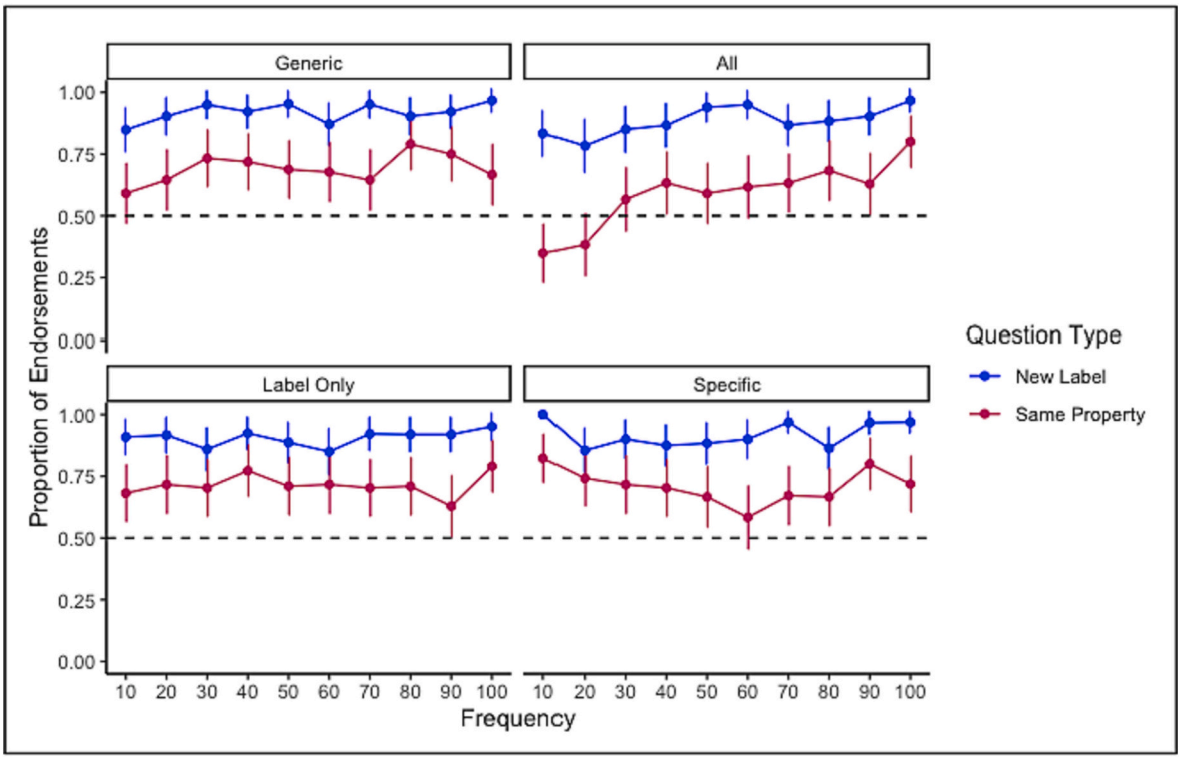

Fig. 4. Proportion of endorsements ("yes" responses) to the New Label and Same Property questions by Observed Frequency in the Label Only, Specific, Generic, and Quantified Conditions. Error bars represent 95% confidence intervals of the mean. Dotted lines represent chance performance (50%). Comparing patterns across conditions, this figure shows that endorsements of claims made about the properties of new objects to "make the machine go" were lowest when the informant had previously made Quantified "all" claims about the properties and the observed frequencies of the property in the sample were low. This contrasted with a general willingness to endorse the informant who had previously made generic claims.

Table 2
Parameter Estimates of the mixed model predicting Knowledge Attributions by condition. Testimony is dummy coded so that the Label Only condition is the reference category.*

Parameter	Estimate	SE	*t (df = 2488)*
Intercept of Label Only condition	**0.92**	0.04	21.24***
Slope of Label Only condition	**0.002**	0.001	3.30**
Intercept difference between Label Only and Specific Conditions	**0.31**	0.06	5.12***
Intercept difference between Label Only and Generic Conditions	**0.02**	0.06	0.32
Intercept difference between Label Only and Quantified ("all") Conditions	**−0.18**	0.06	−2.99**
Slope difference between Label Only and Specific Conditions	**−0.001**	0.001	−0.53
Slope difference between Label Only and Generic Conditions	**−0.003**	0.001	−3.18**
Slope difference between Label Only and Quantified ("all") Conditions	**−0.004**	0.001	3.92***

Adj. $R^2 = 0.22$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

speaker were always as high or higher than baseline, and endorsements of the generic speaker's claims about the causal property of a novel object were steady regardless of observed frequencies (Fig. 5). This is consistent with prior work showing that adults are more likely to endorse the knowledge of, and learn from, speakers who make generic versus specific claims (Butler & Markman, 2014; Cimpian & Park, 2014; Koenig et al., 2015). Altogether, these findings suggest that trust in generic testimony may lead learners to question how well a sample of observations represents the category as a whole and therefore to adjust away from the extremes in their inductive inferences.

Quantified "all" claims ("*All* blickets make the machine go") led to higher prevalence estimates than baseline across the range of observed frequencies. They also led to the greatest degree of skepticism about the knowledge of the informant, and about his subsequent claims to know. But this skepticism did not seem to be based on a straightforward logical relation between the sample and the claim. If participants were treating their own observations as counterevidence against the claim, then any frequency below 100% would have overwhelmingly led to responses that the informant knew "nothing" or "little" about the category (but this result was not obtained), and potentially would also have led participants to reject subsequent claims. Instead, the relation between knowledge attributions and frequency paralleled the results in the Generic condition. This could be due to a tendency to recall quantified claims as generic (Leslie & Gelman, 2012), so that both types of claims were treated similarly. But it could also have been due to a more general level of trust for kind-based claims, especially those stated with confidence, combined perhaps with an uncertainty about probabilistic evidence coming from a small sample of category members.

In both Generic and Quantified ("All") conditions, knowledge attributions predicted prevalence estimates above and beyond the evidence available in the testimonial claims themselves. Because these effects are correlational, our data are consistent with three competing interpretations. One possibility, suggested above, is that participants' beliefs about the knowledge of the speaker influenced prevalence estimates. That is, when the speaker made category-based claims, participants may have weighed those claims more heavily because they believed he was more knowledgeable. Another possibility is that participants' beliefs about the statistical evidence influenced their knowledge ratings. More specifically, participants who were willing to generalize the property widely despite low observed frequencies in the

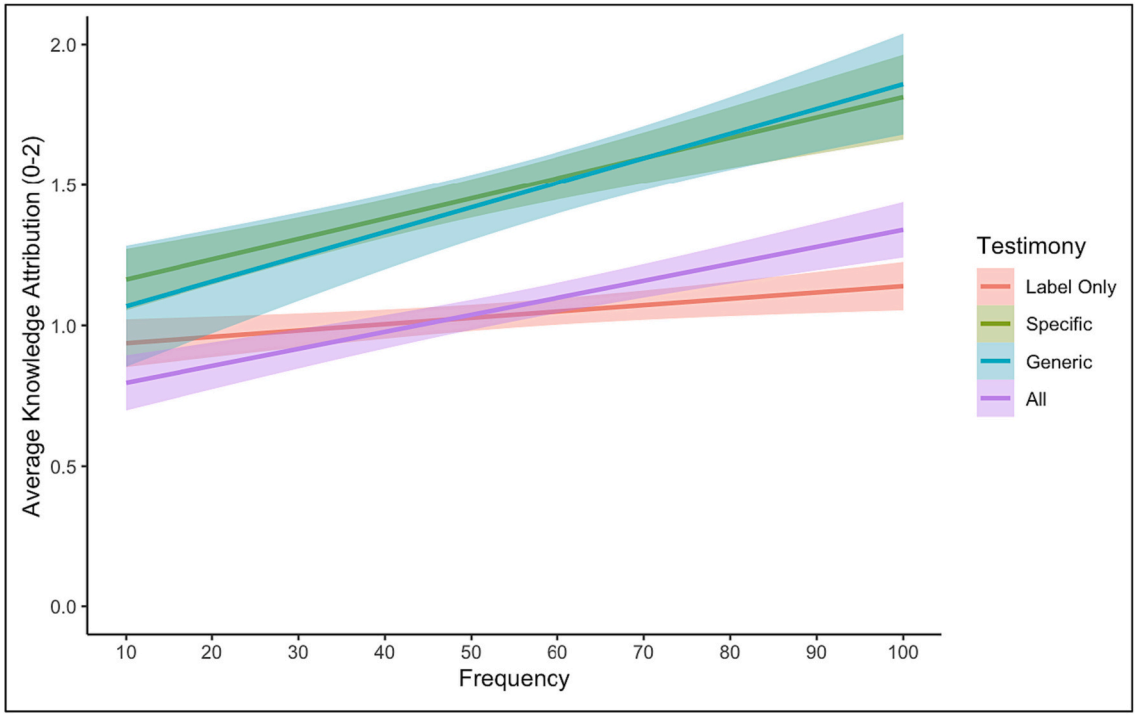

Fig. 5. An illustration of the main effects of Frequency, Testimony, and interaction between the two on participants' knowledge attribution ratings.

sample (those who overestimated prevalence) were thus also more likely to endorse claims that generalized the property ("blickets/all blickets make the machine go"), and this may have resulted in rating the speaker as more knowledgeable. A third possibility is that the correlations we observed reflect some underlying individual difference between participants in how much they weigh their own observations or the language of others in learning something new.

Regardless of the direction of the effects, however, this correlation, taken together with the main effects of both statistical evidence and testimonial language on inductive inferences, suggest that when adult learners have access to both sources of information, they use both sources of evidence in their reasoning. That is, adult learners use information about testimony to verify their observations and use their observations to verify the speakers' claims. Thus, it is both important and revealing that prevalence estimates were more strongly tied to knowledge attributions when participants heard kind-relevant testimony (Quantified or Generic), but not when they heard specific claims or just bare labels. This makes sense when considering which types of claims are easy to verify with one's own observations and which are less so. Thus, our results support a growing number of studies in showing that generic claims are both more likely to be believed (Koenig et al., 2015), and, separately, more likely to "count" as evidence for category learning than other types of claims (Gelman et al., 2010; Rhodes et al., 2012).

2. Study 2

In Study 2, we investigate how testimony and observation interact for young children. Generics abound in child-directed speech (Gelman, Hollander, Star, & Heyman, 2000) as well as children's own speech (Gelman, Goetz, Sarnecka, & Flukes, 2008). By roughly 2.5 years of age, children appropriately produce generics ("Does lions crawl?"; "Boys don't ever be ballet dancers"; Brandone, Cimpian, Leslie & Gelman, 2012; Gelman, Leslie, Gelman, & Leslie, 2019), comprehend generics as kind-referring and distinct from specific reference (Graham, Gelman, & Clarke, 2016), and recall whether information was provided using generic or specific language (Gülgöz & Gelman, 2015). By age 2–3 years, children are exquisitely sensitive to the semantic implications of generics as expressing conceptually central generalizations, and implying that a category is homogeneous, inductively rich, and stable over contexts (Cimpian, 2010; Gelman et al., 2010; Gelman & Brandone, 2010; Graham, Nayer, & Gelman, 2011; Rhodes et al., 2012).

In Study 1, the influence of testimony on adult learners was most apparent for low-frequency observations. Thus, in Study 2 we chose to focus on the 30% condition, in which adult participants consistently overestimated prevalence. Because of the high degree of conflict between the expectations formed by kind-relevant testimony and the low-frequency observations, we believe that this provides the most interesting avenue for understanding how children prioritize one source over the other, and/or integrate the two sources of information.

Children were tested across the same four conditions (Label-only, "All", Generic, Specific) as adults. We used the same videos as in the corresponding conditions of Study 1. The procedure for children was modified in three ways. Two modifications were minor: The questions were worded more simply, in order to obtain property prevalence estimates in a child-friendly way, and additional memory checks were included, to ensure that children understood the format of the questions and were able to track the observed frequencies as they watched the videos. The more substantial modification was that we did not include knowledge questions. Pilot testing revealed that, regardless of condition, nearly all children were overly optimistic about the informant's knowledge of objects and the machine (i.e., they said he knew "a lot"). This did not, however, prevent us from addressing our central question of whether testimony would interact with observations to influence prevalence estimates, and whether it would do so in a way that is similar to or different from adults.

As mentioned above, prior work led us to expect similarities between adult and child learners, but also important differences. One potential developmental difference might be in the power of labels in influencing children's category-based induction. Even in the absence of evidence of a property, labels lead children to make inductive inferences about hidden, non-obvious properties of natural kinds (Gelman & Markman, 1986) and causal properties of novel physical objects such as those in our study (Gopnik & Sobel, 2000). Thus, the presence of a category label

(e.g., "blicket") alone might be enough to lead children to overestimate property prevalence above and beyond the observed frequencies, or even to disregard counterevidence from their own observations (see Schulz, Bonawitz, & Griffiths, 2007, for a similar finding in children's exploratory play). Should this be the case, we may see overestimations in children's prevalence estimates when the category is labeled, even when no claims about the property are made. One exception to this might be specific claims ("*This blicket* makes the machine go"), which imply that the label is not sufficient for property generalization. We tested both of these predictions by focusing our analyses on comparisons between testimony conditions and by directly comparing children's responses to adults'.

Another potential difference might be the influence of testimony on children's memory for observed events, in particular when those events are probabilistic. Observing probabilistic data creates uncertainty, and although children are capable of resolving this uncertainty on their own (Kushnir & Gopnik, 2005; Schulz & Sommerville, 2006), they also rely on testimony more heavily in such cases (Bridgers, Buchsbaum, Seiver, Griffiths, & Gopnik, 2016; McLoughlin, Finiasz, Sobel, & Corriveau, 2021; Plate, Shutts, Cochrane, Green, & Pollak, 2021). Evidence for the influence of testimony on memory is mixed. In daily life, testimony can lead to children's selective remembering of some events over others (Nelson & Fivush, 2004) or even to false memories (Ceci & Bruck, 1993). In word learning, memory and inference are dissociated: children block information from unreliable speakers, but episodic memory for the claims themselves is left intact (Sabbagh & Shafman, 2009). As such, our procedure included two memory questions, and children's responses were compared between testimony conditions.

2.1. Method

2.1.1. Participants

One hundred and twenty-three 4- and 5-year-olds ($M_{age} = 5.06$ years, SD = 0.61 years, range 4.01–5.99 years) participated. Each child was randomly assigned to the Generic condition ($n = 31$), the Specific condition ($n = 29$), the "All" condition ($n = 33$), or the Label Only condition ($n = 30$). Three additional children were tested but excluded from analysis, for not completing the study ($n = 1$) or because they were older siblings outside of the target age range ($n = 2$). An a priori power analysis (ANCOVA, sufficient power for main effects and interactions) determined a sample size of 124 was necessary to obtain 80% power.

2.1.2. Procedure

Children were tested virtually via zoom. An adult parent or guardian was asked to open a Qualtrics survey and share their screen with the experimenter. Adults were also asked to refrain from providing their child with any hints or guidance during the survey and were only allowed to help their child click through the survey and record their answers. All children were able to complete the survey without adult interference.

Children were randomly assigned to one of four Testimony conditions – Generic, Specific, "All", or Label Only. The Observed Frequency of the property in each condition was identical and low – 3 out of 10 objects made the machine go. Children saw two Trials, the first with "blickets" and the second with "midos," as in Study 1.

Children began by completing three training questions. They saw an array of ten items (for example, five cars and five boats) and were asked to select certain items from this array (e.g., "Can you click on the boats"). They were also told to let the experimenter know whenever they were "all done clicking." This allowed children to become familiar with the process of clicking on items on their own screens. All children completed the training questions without error (e.g., clicking on the boats and failing to click on the cars).

After the training phase, children began by watching the video of Zorg and the objects on Trial 1 ("blickets"). As in Experiment 1, Zorg's testimony about the objects varied by condition. In the Generic and "All" conditions, Zorg made a claim about the category (e.g., Generic: "Blickets make the machine go"; "All": "All blickets make the machine go"). In the Specific Condition, Zorg made a claim about one item in the set (e.g. "This blicket makes the machine go"), and in the Label Only Condition Zorg just labeled the items (e.g. "Wow, look at those blickets!"). Immediately after this first presentation of the video, the experimenter asked **Memory Question 1** while showing a still from the movie: "Remember these blickets. These are the ones from the movie, the ones that already went in the machine. Do you remember which ones made the machine go? Can you click on the ones that made the machine go?" (see Fig. 6, left hand side). After this, the movie was played one more time. Showing this question between the two presentations of the video helped to motivate the children to track the objects as they went into the machine (anecdotally, they showed signs of "checking" their memories against the second presentation of the video by vocalizing their recollections and "corrections" to them as the second video played).

Following both video presentations, we asked four questions in a counterbalanced order: (1) For **Memory Question 2** we followed the same form as Memory Question 1. (2) For the **Prevalence Estimate Question**, children were shown an array of 10 objects belonging to the category ("blickets" or "midos" on Trial 1 or 2 respectively) without the informant or machine (see Fig. 6, right hand side). They were told, "These are new ones; we haven't seen them go in the machine yet." Then they were asked, "Which ones will make the machine go? Can you click on the ones that will make the machine go?" (3) For the **Prediction Question**, children were shown a picture of one object in the category ("blicket" or "mido"). They were told, "Okay, here's another one just like the ones you saw before." Then they were asked to label the object, "Do you remember what this is?" and to predict whether it would have the property, "Do you think this blicket [mido] makes the machine go?" (4) For the **Generalization Question**, children were shown a new object (different shape and color) and asked to identify it as belonging to the category or not, "Is this a blicket [mido]?", then asked whether the property would generalize to the new object, "Do you think it makes the machine go?"

Coding. A research assistant who was not informed of the hypotheses or conditions recorded the participants' responses from the Qualtrics survey. Memory questions and Prevalence estimates were coded for the number of objects selected (out of 10). Prediction questions were coded for whether the child said "yes" or "no." Generalization questions were coded for whether the child said the object belonged to the blicket/mido category, and whether the child said "yes" or "no" to the generalization question. A second research assistant spot-checked 10% of the data for errors, and no discrepancies were found.

2.2. Results

2.2.1. Prevalence estimates

To compare children's responses on the prevalence estimates across testimony conditions, we ran a linear mixed model using Testimony (*Generic, Specific, "All", Label-Only*, between subjects), Trial Order (I or II, within subjects), and Age (in months) as predictors. We found main effects of Testimony ($F(3,116) = 4.806$, $p = .003$) and Age ($F(1,116) = 5.06$, $p = .026$). Trial Order was not significant ($p = .711$). Fig. 7 shows the predicted prevalence estimates and 95% confidence intervals from the model. Post-hoc comparisons between conditions revealed that children in the Specific condition chose significantly fewer objects than those in the Label Only and "All" conditions (Fisher's LSD: *Specific-Label Only*: $p = .019$; *Specific-All*: $p < .001$). There were no other significant differences between conditions (Fisher's LSD: all $ps > 0.056$). Overall, older children chose fewer objects than younger children ($r(121) = 0.219$, $p = .016$).

2.2.2. Prevalence estimates: adults and children

We then compared children's responses on the prevalence estimate

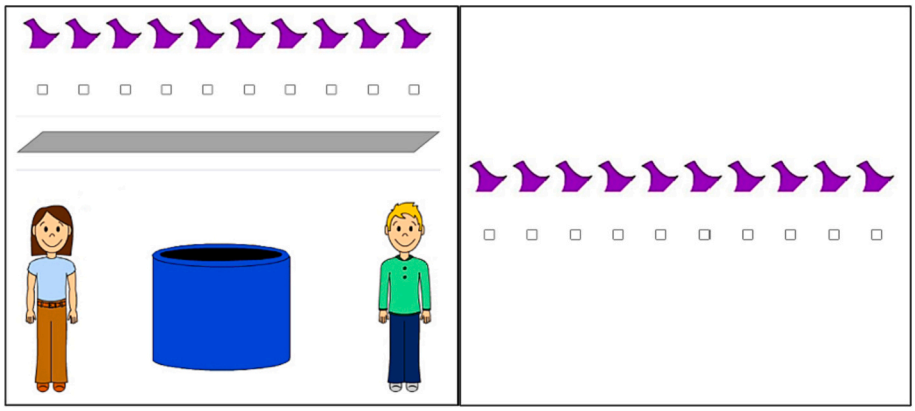

Fig. 6. An example of a Memory (left) and Prevalence (right) Question from Study 2. The memory question showed a still from the video, and children were asked if they "remember the ones that made the machine go" and point to them. The prevalence question showed a new array of objects from the same category and children were asked to point to ones that would make the machine go. The first memory question was asked between the first and second video presentations. The second memory question was counterbalanced with the prevalence, prediction, and generalization questions.

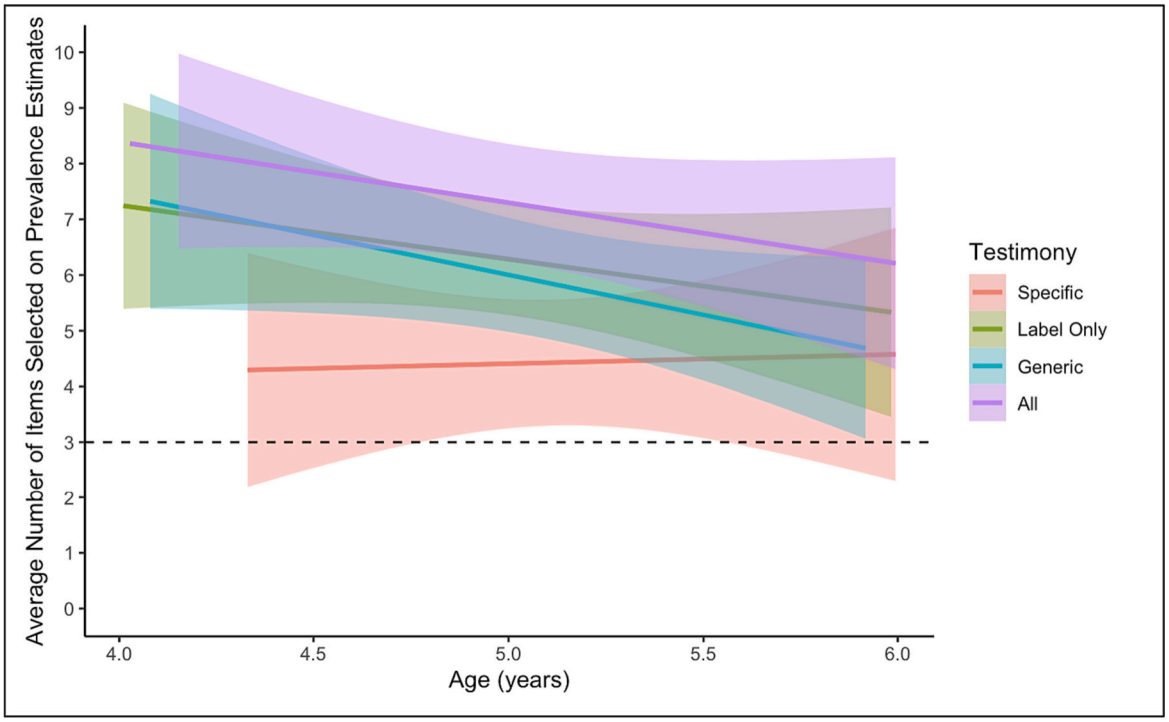

Fig. 7. The model-predicted number of objects children pointed to on the Prevalence Estimates as a function of Age and Testimony. The dotted line indicates the observed frequency (3 objects) of the property in the sample.

questions to adults' responses in the analogous conditions (30% observed frequency, $N = 122$). Because adults' responses were out of 100, we divided by 10 to match the scale used with children. We ran a linear mixed model using Age Group (Adult, Child, between subjects), Testimony (Generic, Specific, "All", Label-Only, between subjects), Trial Order (I or II, within subjects), and the 2-way interaction of Age Group and Testimony as predictors. There were main effects of Testimony ($F(3, 226) = 61.28$, $p < .001$) and of Age Group ($F(1, 226) = 76.06$, $p < .001$). There was no significant Age Group*Testimony interaction ($F(3, 226) = 2.00$, *ns*) and no significant effect of Trial Order ($F(3, 226) = 0.44$, *ns*). Fig. 8 shows the average number of objects selected on the prevalence estimates by Testimony and Age Group. Children selected more objects than adults in all conditions.

We next compared the average responses to the prevalence questions to the observed frequency (3), as a function of age group and testimony conditions. Fig. 8 shows how the mean response in each condition and age group compared to the observed frequency. One-sample *t*-tests show that for children, these averages were significantly above the observed frequency in all four conditions (all $ps < 0.01$). For adults, these averages were significantly different from the observed frequency only in cases where they had heard kind-referring claims (Generic and "All" conditions).

2.2.3. *Memory questions*

To compare children's responses on the memory questions across Testimony conditions, we ran a linear mixed model using Testimony Condition (Generic, Specific, "All", Label-Only, between subjects), Trial Order (I or II, within subjects), Question Order (Memory 1 or Memory 2, within subjects), and Age (in months). There was a main effect of Testimony Condition ($F(3,116) = 7.98$, $p < .001$), and a main effect of Age

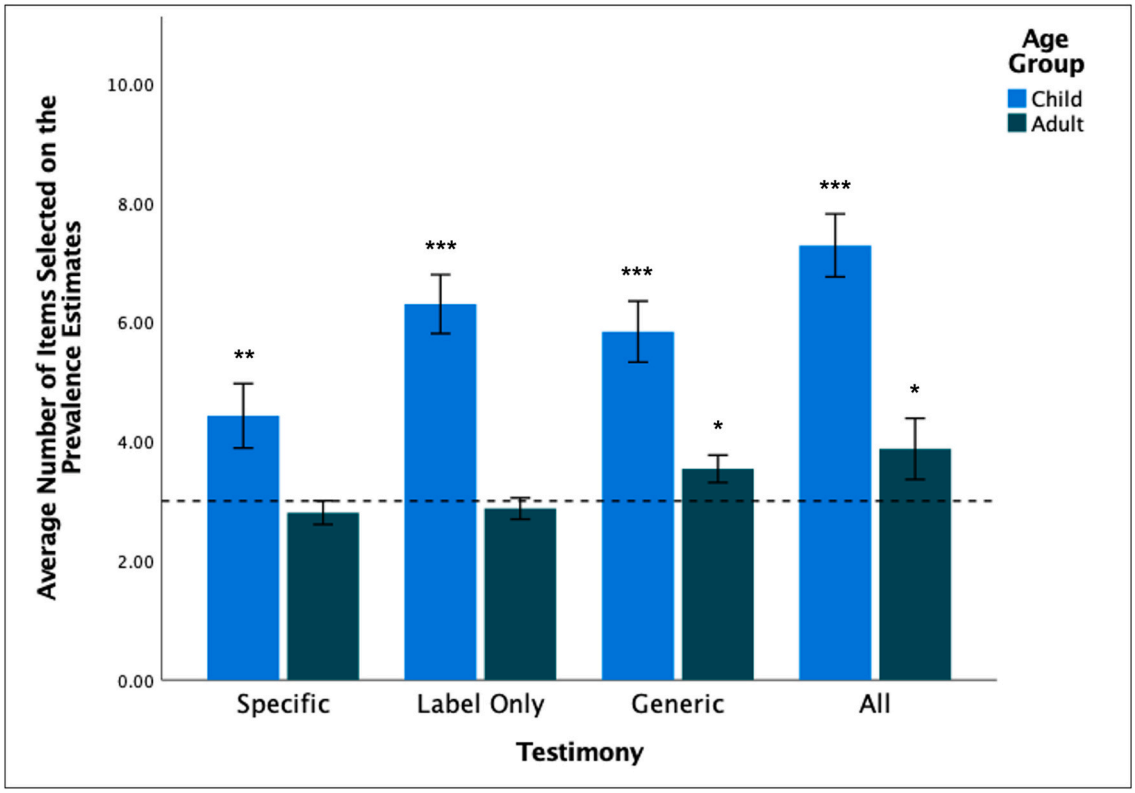

Fig. 8. Average number of objects selected (out of 10) on the prevalence estimate questions by Children and Adults in each Testimony Condition. The dotted line indicates the observed frequency (3 objects) of the property in the sample, asterisks represent one sample *t-test* comparisons to the observed frequency. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

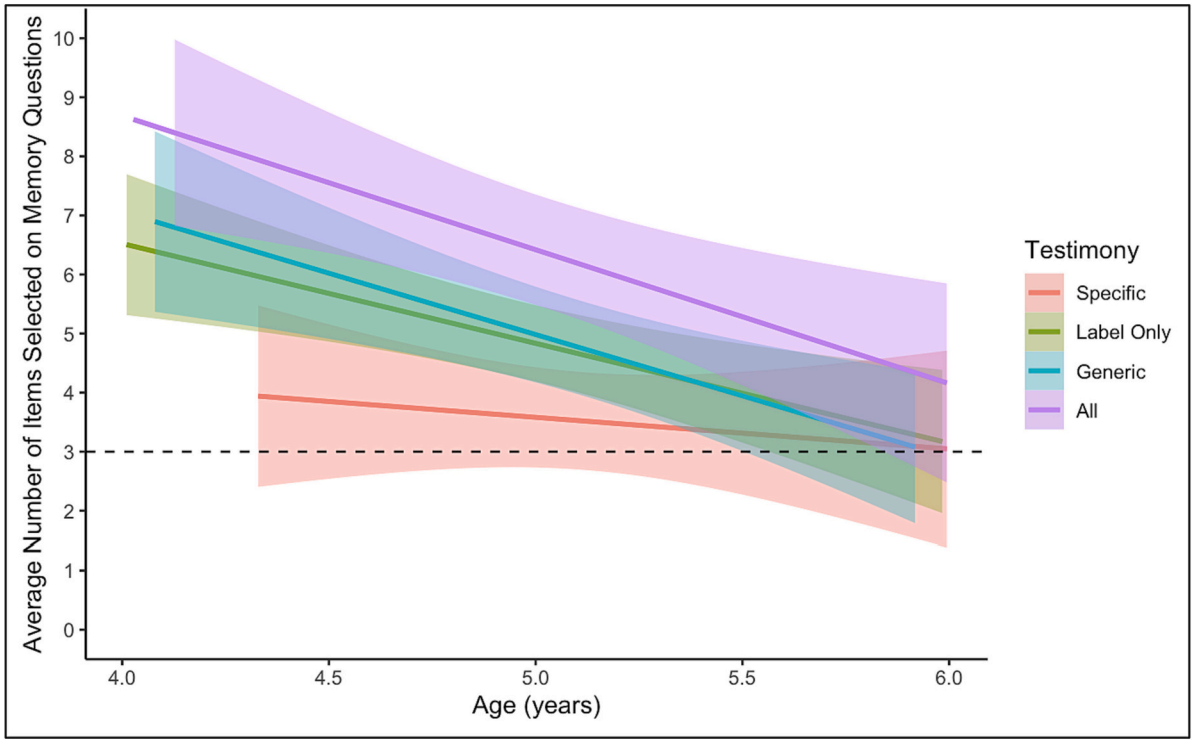

Fig. 9. The model predicting the number of objects children pointed to on the Memory Questions as a function of age and Testimony. The dotted line indicates the observed frequency (3 objects) of the property in the sample.

($F(1,116) = 30.40$, $p < .001$). There were no effects of Trial Order ($p = .701$) or Question Order ($p = .557$). Fig. 9 shows the responses to the Memory questions and 95% confidence intervals from the model. Post-hoc comparisons between conditions revealed that children in the Specific condition chose significantly fewer objects than those in the Generic and "All" conditions (Fisher's LSD: Specific-Generic: $p = .034$; Specific-All: $p < .001$). Conversely, children in the "All" condition chose significantly more objects than those in the other three conditions (Fisher's LSD: All-Generic: $p = .003$; All-Label Only: $p = 001$). There were no other significant differences between conditions (Fisher's LSD: all $ps > 0.071$).

We next analyzed the relation between age and responses to the memory questions. Older children's memories more accurately reflected the observed frequencies than younger children's memories ($r(122) = 0.44$, $p < .001$). This pattern of results remains when looking at conditions where children heard either quantified or kind-referring testimony (*Label Only:* $r(30) = -0.54$, $p = .002$*; Generic* $r(31) = -0.53$, $p = 002$*; All:* $r(32) = -0.51$, $p = .003$). For children in the Specific condition, there was no relation between memory accuracy and age ($r(29) = -0.13$, *ns*) and children's responses to the memory question in this condition were accurate (one sample *t*: $t(29) = 1.36$, *ns*).

2.2.4. Prediction and generalization questions

Children's responses to the Prediction Questions did not differ across trials (McNemar's test, *ns*) and were overwhelmingly positive (81/90 children said "yes" to the prediction question on both trials). Children's responses to the Generalization Question also did not differ across trials (McNemar's test, *ns*). However, their responses to the Generalization Question depended on whether they thought the new object was a member of the category. On Trial 1, out of the 77 children who said the new object was *not* a blicket, 72 of the 77 (94%) predicted it would *not* make the machine go. Of the 9 children who said the new object *was* a blicket, all 9 of them (100%) predicted it *would* make the machine go (Fisher's exact test, $p < .001$). Similarly, on Trial 2, 67/80 (84%) children who said the new object was not a mido predicted it would not make the machine go, and 9/9 children (100%) who said the new object was a mido predicted it would make the machine go (Fisher's exact test, $p < .001$). Children's responses to the Generalization Question did not depend on condition. A one-way ANOVA on the number of times children generalized the property to the new object by condition was not significant ($F(2,169) = 1.85$, *ns*).

2.3. Discussion

Study 2 investigated the role of testimony and observations on category-based inferences in 4- to 5-year-old children. Children heard a testimonial claim, observed a property that was true of 30% of a set of novel objects, and then were asked to estimate the prevalence of the property in 10 new category members, to remember what they saw, and to predict the properties of two new items. Much as for adults in Study 1, kind-referring testimony – testimony that included either a quantified "all" or a generic claim – led children to overestimate the prevalence of the causal property in new instances. Also, like adults, children's estimates after hearing testimony about one specific category member were closer to the exact probabilities that they observed. This general contrast between estimates following kind-referring versus specific language held despite age-related changes in children's probability estimates.

There were also interesting differences between children's and adults' responses to the prevalence estimate questions. Children's estimates were higher than adults' on average across all conditions, with this tendency to overestimate decreasing with age. Another difference was that children (especially younger children) overestimated prevalence even when they were given just a category label and *no* property information. Finally, and unexpectedly, testimony influenced younger children's reporting of observed events (Memory questions) and did so differentially depending on what kind of testimony they heard. Specifically, the youngest children were more likely to report having *seen* more objects with the causal property when they heard testimony that implied high prevalence ("all" or generic claims) than when they heard testimony that pointed to the property in one object only. Memories after hearing the category label alone (Label Only condition) fell in between.

One explanation for children's overestimations may be rooted in children's deference to testimony. There is a large body of work showing that children are generally trusting of adult testimony, even when their claims are counterintuitive, unexpected, or even disproven (Jaswal, 2010; Jaswal, Croft, Setia, & Cole, 2010; Koenig & Harris, 2005a, 2005b). Further, this may be particularly true of younger children. Younger children's responses were higher than older children's, and the number of younger children who selected all ten objects when testimony included an "all" quantifier was significantly higher than the number of older children and was condition dependent. Children's overestimation following labeled testimony is also consistent with research showing that children view labels as kind-referring (Gelman & Brandone, 2010; Waxman & Markow, 1995) and that labeling objects encourages children to make inductive inferences about non-obvious "insides" and causal properties of category members (Gelman & Markman, 1986; Gopnik & Sobel, 2000; Graham, Kilbreath, & Welder, 2004; Taborda-Osorio & Cheries, 2018; Waxman & Braun, 2005). Finally, it should be noted that these two interpretations –that children are particularly influenced by testimony and that labels are special for children– are not mutually exclusive. Rather, both may have worked in tandem, leading to children's consistent overestimations on the prevalence questions.

The condition differences in younger children's responses to the memory questions suggest another mechanism by which testimony can influence children's category learning. There are several reasons why children might have reported remembering more objects having the property than what they saw. First, we rule out difficulties with attending to the relevant information, since children across the age range were accurate in the Specific condition. We are then left with two explanations – source monitoring errors due to interference of testimony (Drummey & Newcombe, 2002; Mahr & Csibra, 2021; Mahr, Mascaro, Mercier, & Csibra, 2021), or pragmatic errors in interpreting the memory question itself. Either way, these findings are intriguing, and expand on prior work showing that, when evidence is uncertain, children defer more to testimony than their own observations (Bridgers et al., 2016; McLoughlin et al., 2021). Our results extend this work by showing that testimony, especially kind-referring testimony, changes the way that children remember their own observations.

3. General discussion

We learn about categories in many different ways. One source of evidence is testimony – we can rely on others' kind-based claims to make inferences about members of a category. Another source of evidence is observation. By sampling from a category, we can observe the common characteristics of category members and use this information to make inductive generalizations about other category members. The current study addresses how these sources of evidence work together when testimony and observation conflict to varying degrees. Our findings show that rather than relying solely on testimony or solely on observation, learners of all ages integrated both sources of information to make inferences, attribute knowledge (adults in Study 1), and remember their own observations (children in Study 2). When these two sources of information conflict, the degree to which learners prioritize one source over the other is dependent on the kind of testimony they receive, the degree of trust or skepticism in a speaker's knowledge, and – for children – their memories of what they had seen.

The degree to which adult learners in our study estimated the prevalence of the causal property in the category was systematically linked to their observations – lower frequency observations led to lower estimates, higher frequency observations led to higher prevalence

estimates. The degree to which learners deviated from the prevalence of the observations was also linked to the kind of testimony they received. Generic and quantified "all" claims about a property of the category led adults and children to overestimate that property's prevalence. For adults, this was particularly true when the observed frequency was low. Notably, generic testimony had the effect of "flattening" adults' prevalence estimates: participants overestimated prevalence at low observed frequencies and underestimated at high ones. These results extend prior work on the semantic implications of generic claims – namely that they imply high prevalence, but also cannot be mapped onto a particular frequency or probability (Butler & Markman, 2014; Cimpian & Markman, 2009; Prasada, 2000; Tessler et al., 2020). When the category was labeled but no information about the property was given, children, but not adults, overestimated prevalence. After hearing specific claims, both children's and adults' estimates were more closely matched to their observations.

Questions remain about the exact mechanisms driving the interaction between observation and testimony. One possibility is that generic claims give the impression of being more informative than small samples, leading learners to question the representativeness of their own observations. Another possibility, not mutually exclusive, is that the semantics of generic testimony productively combine with a general trust of confident generic speakers (Koenig & Harris, 2005a, 2005b). Indeed, our own results support this second possibility, as adults in Study 1 rated generic speakers as most knowledgeable, and individual learners who thought generic speakers were more knowledgeable were also more likely to overestimate prevalence above and beyond what they observed. On the other hand, when observed frequencies were high, adult learners had a slight tendency to undergeneralize the property to a new set of objects after hearing generic testimony (rather than precisely quantified "all" claims).

Learners are more likely to endorse claims from speakers whom they believe are knowledgeable. This is evidenced by the fact that adults in our study who attributed more knowledge to the informant also generalized the property at higher rates, suggesting that more credulous adults may have favored testimony over their own observations. Further, adults' knowledge attributions were far from floor levels, even in cases where the mismatch between testimony and observation were greatest and skepticism would theoretically be high. We suggest that this is related to the fact that the speaker stated his claims with confidence across all conditions. This is consistent with prior work showing that confident claims are more likely to be endorsed (Birch, Severson, & Baimel, 2020; Koenig & Harris, 2005a, 2005b; Sobel & Kushnir, 2013). However, adults' knowledge attributions were not uniform – the degree to which they attributed knowledge also depended on the type of testimony they received. When testimony was easily disproven by observation (as in the "all" condition), knowledge attributions were lowest. This was especially true when the frequency of the observed property was lower, and the mismatch between testimony and observation was greatest. However, when testimony contained a generic statement, adults rated the speaker as most knowledgeable, which is especially notable following low-frequency observations. We suggest that this is because generics allow for exceptions (Hollander et al., 2002; Leslie, 2008), meaning that participants were not able to use their observations to reject the speaker's claims. Together these findings show the unique properties of generics when learning from testimony: Generic claims are harder to falsify, and as a result, they may be more compelling than other types of testimony.

Although children were not explicitly asked to attribute knowledge to the speaker, their responses to the memory questions suggest that they trusted what they heard. Overall, children were more credulous than adults. For example, following kind-referring testimony and testimony that only provided a label and no property information, they recalled the prevalence at rates that were higher than what they had observed. Notably, older children's responses to the memory questions were closer to what they had actually observed, whereas younger children overestimated at higher rates. This finding indicates that younger children may have been more trusting of the informant, leading them to favor what they had been told over their own observations. Older children, on the other hand, were less credulous and more likely to accurately recall what they had seen. This extends prior work showing that children are less deferential to testimony as they get older (Chan & Tardif, 2013; Jaswal, 2004; Jaswal et al., 2010; Ma & Ganea, 2010; Mills, 2013). Finally, children's responses to the memory questions and prevalence estimates were related, as higher responses on the memory questions also meant higher responses on the prevalence estimates (see supplement). This is consistent with the idea that endorsement of the speaker's claims led to inflated estimates – that is, generalizing the property at higher rates than were observed.

One intriguing difference between adult and child participants was the role of testimony containing only a label (and no other information about object properties) in category learning. For children, testimony containing only a label was a special case: children who heard testimony with only a label responded in a similar way to those who had heard generic testimony. One possible interpretation of this finding is that, for children, the mere existence of a shared category label implied that objects in the category had shared traits. This possibility is supported by existing work on labeled categories showing that children infer shared properties amongst objects that share a label (Gelman, 2003; Gopnik & Sobel, 2000; Nazzi & Gopnik, 2000). Moreover, there is some evidence for the generative power of shared labels in responses to the prediction and generalization questions. Children who thought the novel object was a blicket also believed it would activate the machine, whereas those who did not think it was a blicket did not believe it would activate the machine. An open question that remains is why adults who heard testimony containing only a label did not overgeneralize on the prevalence estimate questions, given that they also make similar inferences from category labels (Deng & Sloutsky, 2013).

Finally, it is important to note that our effects are most apparent when observed frequencies were low. We believe that this pattern of results highlights the role of uncertainty when learning form testimony and observation. A learner may feel more of this uncertainty when there is more conflict between the two sources of information. Indeed, probabilistic data creates uncertainty especially when paired with generic or "all"-quantified testimony that implies high prevalence (McLoughlin et al., 2021; Plate et al., 2021). Thus, when kind-referring testimony was followed by low-frequency observations, participants may have felt more uncertain, leading them to prioritize confidently stated testimony over their own observations (Tenney, Small, Kondrad, Jaswal, & Spellman, 2011; also see Plate et al., 2021). Indeed, the contrast between the lower and higher frequencies for adults who heard "all" quantified testimony demonstrates that, when uncertainty is higher, learners may be more inclined to prioritize testimony in their inductive generalizations. For instance, following "all"-quantified testimony, adults overgeneralized at the lower ends, when there was a higher degree of contrast, even though their attributions of knowledge were lower. This suggests that, although they may have been skeptical of the informant, his testimony still led them to question whether their observations were actually representative of the entire sample. Similarly, children have been shown to prefer confident speakers (Brosseau-Liard, Cassels, & Birch, 2014; Jaswal & Malone, 2007; Sabbagh & Baldwin, 2001) and they rely more heavily on confidently stated testimony when observing probabilistic data (Bridgers et al., 2016; McLoughlin et al., 2021; Tenney et al., 2011). Future work on children's generalizations when frequencies are high (and observations are better calibrated to "all"-quantified and generic testimony), as well as including more explicit measures of certainty for both adult and child learners may help tease apart the role that a learner's certainty plays when they are receiving information from both testimony and observation.

We believe these findings have implications for category learning outside the lab, potentially shedding light on a mechanism by which testimony about social stereotypes is integrated with observation. For

instance, hearing a generic statement such as "Girls are bad at science" has negative impacts for girls' science learning (Bian, Leslie, & Cimpian, 2017; Catsambis, 1994; Herbert & Stipek, 2005; Zhao, Seibert, & Hills, 2005), despite initially equal STEM aptitude for boys and girls (Else-Quest, Hyde, & Linn, 2010) and the many observable instances of successful women in STEM. In contrast, specific testimony may have a protective effect by encouraging learners to limit their extensions of a property beyond the specific referent of that statement (see also Foster-Hanson, Roberts, Gelman, & Rhodes, 2021; Gelman et al., 2010; Moty & Rhodes, 2021; Rhodes et al., 2012; Rhodes, Leslie, Bianchi, & Chalik, 2018; Roberts, Ho, Rhodes, & Gelman, 2017).

Similarly, the current methodology may be useful for understanding the transmission of information about other kinds of knowledge. Scientific learning often makes use of both testimony and observation to demonstrate phenomena. This is particularly true for non-obvious or non-visible causal relationships (e.g., fire needs oxygen to burn), and for causal relationships that do not occur 100% of the time (e.g., smoking causes lung cancer). For instance, work on children's understanding of unexpected causes (e.g., being scared causes stomach aches) has explored how children's expectations interact with the evidence they observe (Schulz et al., 2007). We may expect similar results if children's initial expectations are formed by receiving testimony. Similarly, work on scientific exploration has shown how children's conversations with their parents (testimony) changes how they explore and generate evidence in a causal system (observation) (Medina & Sobel, 2020; Sobel & Stricker, 2022; Willard et al., 2019). Together, these studies suggest that across domains (social, biological, physical) we may expect to see a similar effect of testimony on the way observations are used to learn. We leave future work to expand the current investigation to these other domains.

The current studies present a number of limitations and open questions. The studies were conducted in English, and as a result, caution should be taken when generalizing these findings to speakers of other languages. For instance, prior research has shown that generic knowledge is expressed differently across languages, and the use of generic noun phrases may follow different developmental timelines (Gelman & Tardif, 1998; Goldin-Meadow, 2005; Krifka et al., 1995; Mannheim, Gelman, Escalante, Huayhua, & Puma, 2010; Tardif, 1996). A second limitation is that we did not capture children's explicit knowledge attributions. Prior work has highlighted the role of trust and attribution of epistemic competence when children are learning from testimony (Harris, 2012; Harris & Koenig, 2006; Koenig et al., 2015; Sobel & Kushnir, 2013). As such, future work should use more explicit measures to explore the role of knowledge attribution in children's category learning.

Additionally, two open questions remain about the timing and order in which testimony and observation are presented. In the present studies, testimony was immediately followed by observation. However, the order can be reversed, and indeed learners may observe something long before they receive any testimony about it. If learners have a chance to form their own beliefs before hearing testimony, they may be more inclined to prioritize their observations than when testimony comes first. Conversely, if testimony appears first and there is a longer delay between testimony and observation, this may strengthen the role of testimony. Future research should examine the order and timing of testimony versus observation.

Additionally, in future work it would be valuable to see what kinds of information participants seek, when given an opportunity. The current study provided a limited set of observations and controlled the frequency of the causal property within this set. However, different results may arise if participants were given the opportunity to test the devices themselves. We suspect that adult and child learners may be more inclined to endorse their own self-generated observations than they would in the current study, as involvement in the information-gathering process may make for more compelling evidence than observation alone (Kushnir & Gopnik, 2005; Markant & Gureckis, 2014; Sobel & Letourneau, 2018). Another interesting avenue would be to allow participants to seek additional testimony regarding the original set of objects, from the original speaker or others. In doing so, we may be able to answer additional questions about whether or not an individual has accepted a speaker's claims. An additional possibility would be to give a child a choice between the two kinds of information (direct observations vs. testimony). We leave future work to explore these questions further.

Both testimony and observation provide a wealth of information beyond the features we have examined, that inform learners' inferences. In deciding which testimony to accept, learners track a speaker's expertise and prior accuracy (Hermes et al., 2020; Koenig & Harris, 2005a, 2005b; Koenig & Woodward, 2010; Kuzyk, Grossman, & Poulin-Dubois, 2020; Sobel & Kushnir, 2013). Children and adults also use information about an informant's social standing (Bernard, Proust, & Clement, 2015; Landrum, Mills, & Johnston, 2013; MacDonald, Schug, Chase, & Barth, 2013; Reyes-Jaquez & Echols, 2013) and physical traits (Bascandziev & Harris, 2014; Todorov, Pakrashi, & Oosterhof, 2009) when deciding which testimony to believe. Observation, too, provides learners with ample evidence to discern whether a property is kind-relevant. For instance, learners account for both typicality and representativeness of a sample when deciding which observed properties to generalize (Gweon, Tenenbaum, & Schulz, 2010; Lo, Sides, Rozelle, & Osherson, 2002; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Shipley & Shepperson, 2006). They also make inferences based on the method by which a sample was obtained (Foster-Hanson, Moty, Cardarelli, Ocampo, & Rhodes, 2020; Xu & Denison, 2009). This leaves open questions about how people integrate different kinds of testimony (from different kinds of informants) with different kinds of statistical evidence.

Learners of all ages are often monitoring these features when appraising evidence from both testimony and observation. In doing so, they are making evidence-based decisions about which claims to accept and which observations to generalize. The current studies have investigated a subset of these features, showing that learners use both the nature of the testimony and the frequency of their observations to infer when a property is kind-relevant. We leave future work to examine the influence that additional factors have on the way that learners integrate both sources of evidence when learning about categories.

Data availability

Included in attached files

Acknowledgments

We would like to thank Natalie Davidson, Dana Karami, Maria Lee, Erin Powder, Janay Saunders, Esha Sheth, and Alexandra Was for help creating stimuli, collecting data, and coding. We also thank Alex Was for her help with piloting an earlier procedure. This research was supported by funding from NSF (DLS #1023179) to T.K. and NICHD grant HD-36043 to S.G.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105707.

References

Bascandziev, I., & Harris, P. L. (2014). In beauty we trust: Children prefer information from more attractive informants. *British Journal of Developmental Psychology, 32*(1), 94–99.

Bernard, S., Proust, J., & Clement, F. (2015). Four- to six-year-old children's sensitivity to reliability versus consensus in the endorsement of object labels. *Child Development, 86*(4), 1112–1124.

Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science, 355*(6323), 389–391.

Birch, S. A., Severson, R. L., & Baimel, A. (2020). Children's understanding of when a person's confidence and hesitancy is a cue to their credibility. *PLoS One, 15*(1), Article e0227026.
Brandone, A. C., Cimpian, A., Leslie, S., & Gelman, S. A. (2012). Do lions have manes? For children, generics are about kinds rather than quantities. *Child Development, 83*(2), 423–433.
Brandone, A. C., Gelman, S. A., & Hedglen, J. (2015). Children's developing intuitions about the truth conditions and implications of novel generics versus quantified statements. *Cognitive Science, 39*(4), 711–738.
Bridgers, S., Buchsbaum, D., Seiver, E., Griffiths, T. L., & Gopnik, A. (2016). Children's causal inferences from conflicting testimony and observations. *Developmental Psychology, 52*(1), 9–18.
Brosseau-Liard, P., Cassels, T., & Birch, S. (2014). You seem certain but you were wrong before: Developmental change in preschoolers' relative trust in accurate versus confident speakers. *PLoS One, 9*(9), Article e108308.
Butler, L. P., & Markman, E. M. (2014). Preschoolers use pedagogical cues to guide radical reorganization of category knowledge. *Cognition, 130*(1), 116–127.
Catsambis, S. (1994). The path to math: Gender and racial-ethnic differences in mathematics participation from middle school to high school. *Sociology of Education*, 199–215.
Ceci, S. J., & Bruck, M. (1993). Suggestibility of the child witness: A historical review and synthesis. *Psychological Bulletin, 113*(3), 403–439.
Cella, F., Marchak, K. A., Bianchi, C., & Gelman, S. A. (2022). Generic language for social and animal kinds: An examination of the asymmetry between acceptance and inferences. *Cognitive Science, 46*(12), Article e13209.
Chan, C. C., & Tardif, T. (2013). Knowing better: The role of prior knowledge and culture in trust in testimony. *Developmental Psychology, 49*(3), 591–601.
Cimpian, A. (2010). The impact of generic language about ability on children's achievement motivation. *Developmental Psychology, 46*(5), 1333–1340.
Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science, 34*(8), 1452–1482.
Cimpian, A., Gelman, S. A., & Brandone, A. C. (2010). Theory-based considerations influence the interpretation of generic sentences. *Language and Cognitive Processes, 25*(2), 261–276.
Cimpian, A., & Markman, E. M. (2009). Information learned from generic language becomes central to children's biological concepts: Evidence from their open-ended explanations. *Cognition, 113*(1), 14–25.
Cimpian, A., & Park, J. J. (2014). Tell me about pangolins! Evidence that children are motivated to learn about kinds. *Journal of Experimental Psychology, 143*(1), 46–55.
Deng, W., & Sloutsky, V. M. (2013). The role of linguistic labels in inductive generalization. *Journal of Experimental Child Psychology, 114*(3), 432–455.
Denison, S., Trikutam, P., & Xu, F. (2014). Probability versus representativeness in infancy: Can infants use naïve physics to adjust population base rates in probabilistic inference? *Developmental Psychology, 50*(8), 2009–2019.
Drummey, A. B., & Newcombe, N. S. (2002). Developmental changes in source memory. *Developmental Science, 5*(4), 502–513.
Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103–127.
Foster-Hanson, E., Moty, K., Cardarelli, A., Ocampo, J. D., & Rhodes, M. (2020). Developmental changes in strategies for gathering evidence about biological kinds. *Cognitive Science, 44*(5), Article e12837.
Foster-Hanson, E., Roberts, S. O., Gelman, S. A., & Rhodes, M. (2021). Categories convey prescriptive information across domains and development. *Journal of Experimental Child Psychology, 212*, Article 105231.
Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York, NY: Oxford University Press.
Gelman, S. A. (2023). Looking beyond the obvious. *American Psychologist, 78*(5), 667–677.
Gelman, S. A., & Bloom, P. (2007). Developmental changes in the understanding of generics. *Cognition, 105*(1), 166–183.
Gelman, S. A., & Brandone, A. (2010). Fast-mapping placeholders: Using words to talk about kinds. *Language Learning and Development, 6*(3), 223–240.
Gelman, S. A., Coley, J. D., Rosengren, K. S., Hartman, E., Pappas, A., & Keil, F. C. (1998). Beyond labeling: The role of maternal input in the acquisition of richly structured categories. *Monographs of the Society for Research in Child Development*, i–157.
Gelman, S. A., Goetz, P. J., Sarnecka, B. S., & Flukes, J. (2008). Generic language in parent child conversations. *Language Learning and Development, 4*(1), 1–31.
Gelman, S. A., Hollander, M., Star, J., & Heyman, G. D. (2000). The role of language in the construction of kinds. *Psychology of Learning and Motivation, 39*, 201–263.
Gelman, S. A., Leslie, S. J., Gelman, R., & Leslie, A. (2019). Do children recall numbers as generic? A strong test of the generics-as-default hypothesis. *Language Learning and Development, 15*(3), 217–231.
Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*(3), 183–209.
Gelman, S. A., Star, J. R., & Flukes, J. (2002). Children's use of generics in inductive inferences. *Journal of Cognition and Development, 3*(2), 179–199.
Gelman, S. A., & Tardif, T. Z. (1998). A cross-linguistic comparison of generic noun phrases in English and mandarin. *Cognition, 66*(3), 215–248.
Gelman, S. A., Ware, E., & Kleinberg, F. (2010). Effects of generic language on category content and structure. *Cognitive Psychology, 61*(3), 273–301.
Goldin-Meadow, S. (2005). What language creation in the manual modality tells us about the foundations of language. *Linguistic Review, 22*(2–4), 199–225.
Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71*(5), 1205–1222.
Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*(6), 1085–1108.
Graham, S. A., Gelman, S. A., & Clarke, J. (2016). Generics license 30-month-olds' inferences about the atypical properties of novel kinds. *Developmental Psychology, 52*(9), 1353–1362.
Graham, S. A., Kilbreath, C. S., & Welder, A. N. (2004). Thirteen-month-olds rely on shared labels and shape similarity for inductive inferences. *Child Development, 75*(2), 409–427.
Graham, S. A., Nayer, S. L., & Gelman, S. A. (2011). Two-year-olds use the generic/non-generic distinction to guide their inferences about novel kinds. *Child Development, 82*(2), 493–507.
Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science, 35*(8), 1407–1455.
Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116*(4), 661–716.
Gülgöz, S., & Gelman, S. A. (2015). Children's recall of generic and specific labels regarding animals and people. *Cognitive Development, 33*, 84–98.
Gweon, H., Tenenbaum, J., & 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.
Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press.
Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1567), 1179–1187.
Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development, 77*(3), 505–524.
Herbert, J., & Stipek, D. (2005). The emergence of gender differences in children's perceptions of their academic competence. *Journal of Applied Developmental Psychology, 26*(3), 276–295.
Hermansen, T. K., Ronfard, S., Harris, P. L., Pons, F., & Zambrana, I. M. (2021). Young children update their trust in an informant's claim when experience tells them otherwise. *Journal of Experimental Child Psychology, 205*, Article 105063.
Hermes, J., Brugger, F., Illner, T., Plate, A., Rakoczy, H., & Behne, T. (2020). *Selective trust in young children and distracted adults: halo-effects outweigh rational choices*.
Hoicka, E., Saul, J., Prouten, E., Whitehead, L., & Sterken, R. (2021). Language signaling high proportions and generics lead to generalizing, but not essentializing, for novel social kinds. *Cognitive Science, 45*(11), Article e13051.
Hollander, M. A., Gelman, S. A., & Star, J. (2002). Children's interpretation of generic noun phrases. *Developmental Psychology, 38*(6), 883–894.
Jaswal, V. K. (2004). Don't believe everything you hear: Preschoolers' sensitivity to speaker intent in category induction. *Child Development, 75*(6), 1871–1885.
Jaswal, V. K. (2010). Believing what you're told: Young children's trust in unexpected testimony about the physical world. *Cognitive Psychology, 61*(3), 248–272.
Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science, 21*(10), 1541–1547.
Jaswal, V. K., & Malone, L. S. (2007). Turning believers into skeptics: 3-year-olds' sensitivity to cues to speaker credibility. *Journal of Cognition and Development, 8*(3), 263–283.
Keil, F. C. (2010). The feasibility of folk science. *Cognitive Science, 34*(5), 826–862.
Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116*(1), 20–58.
Kimura, K., & Gopnik, A. (2019). Rational higher-order belief revision in young children. *Child Development, 90*(1), 91–97.
Koenig, M. A., Cole, C. A., Meyer, M., Ridge, K. E., Kushnir, T., & Gelman, S. A. (2015). Reasoning about knowledge: Children's evaluations of generality and verifiability. *Cognitive Psychology, 83*, 22–39.
Koenig, M. A., & Harris, P. L. (2005a). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development, 76*(6), 1261–1277.
Koenig, M. A., & Harris, P. L. (2005b). The role of social cognition in early trust. *Trends in Cognitive Sciences, 9*(10), 457–459.
Koenig, M. A., & Woodward, A. L. (2010). Sensitivity of 24-month-olds to the prior inaccuracy of the source: Possible mechanisms. *Developmental Psychology, 46*(4), 815.
Krifka, M., Pelletier, F. J., Carlson, G., Ter Meulen, A., Chierchia, G., & Link, G. (1995). *Genericity: An introduction*.
Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science, 16*(9), 678–683.
Kushnir, T., Vredenburgh, C., & Schneider, L. A. (2013). "Who can help me fix this toy?" the distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. *Developmental Psychology, 49*(3), 446.
Kuzyk, O., Grossman, S., & Poulin-Dubois, D. (2020). Knowing who knows: Metacognitive and causal learning abilities guide infants' selective social learning. *Developmental Science, 23*(3), Article e12904.
Landrum, A. R., Mills, C. M., & Johnston, A. M. (2013). When do children trust the expert? Benevolence information influences children's trust more than expertise. *Developmental Science, 16*(4), 622–638.
Leslie, S., & Gelman, S. A. (2012). Quantified statements are recalled as generics: Evidence from preschool children and adults. *Cognitive Psychology, 64*(3), 186–214.
Leslie, S. J. (2008). Generics: Cognition and acquisition. *Philosophical Review, 117*(1), 1–47.

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgment. *Cognitive Science, 26*(2), 181–206.
Ma, L., & Ganea, P. A. (2010). Dealing with conflicting information: Young children's reliance on what they see versus what they are told. *Developmental Science, 13*(1), 151–160.
MacDonald, K., Schug, M., Chase, E., & Barth, H. (2013). My people, right or wrong? Minimal group membership disrupts preschoolers' selective trust. *Cognitive Development, 28*(3), 247–259.
Mahr, J. B., & Csibra, G. (2021). The effect of source claims on statement believability and speaker accountability. *Memory & Cognition, 49*(8), 1505–1525.
Mahr, J. B., Mascaro, O., Mercier, H., & Csibra, G. (2021). The effect of disagreement on children's source memory performance. *PLoS One, 16*(4), Article e0249958.
Mannheim, B., Gelman, S. A., Escalante, C., Huayhua, M., & Puma, R. (2010). A developmental analysis of generic nouns in southern Peruvian Quechua. *Language Learning and Development, 7*(1), 1–23.
Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General, 143*(1), 94–122.
McLoughlin, N., Finiasz, Z., Sobel, D. M., & Corriveau, K. H. (2021). Children's developing capacity to calibrate the verbal testimony of others with observed evidence when inferring causal relations. *Journal of Experimental Child Psychology, 210*, Article 105183.
Medina, C., & Sobel, D. M. (2020). Caregiver–child interaction influences causal learning and engagement during structured play. *Journal of Experimental Child Psychology, 189*, Article 104678.
Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. *Developmental Psychology, 49*(3), 404–418.
Moty, K., & Rhodes, M. (2021). The unintended consequences of the things we say: What generic statements communicate to children about unmentioned categories. *Psychological Science, 32*(2), 189–203.
Nazzi, T., & Gopnik, A. (2000). A shift in children's use of perceptual and causal cues to categorization. *Developmental Science, 3*(4), 389–396.
Nelson, K., & Fivush, R. (2004). The emergence of autobiographical memory: A social cultural developmental theory. *Psychological Review, 111*(2), 486–511.
Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition, 78*(2), 165–188.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.
Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*(2), 185–200.
Plate, R. C., Shutts, K., Cochrane, A., Green, C. S., & Pollak, S. D. (2021). Testimony bias lingers across development under uncertainty. *Developmental Psychology, 57*(12), 2150.
Prasada, S. (2000). Acquiring generic knowledge. *Trends in Cognitive Sciences, 4*(2), 66–72.
Reyes-Jaquez, B., & Echols, C. H. (2013). Developmental differences in the relative weighing of informants' social attributes. *Developmental Psychology, 49*(3), 602–613.
Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition, 108*(2), 543–556.
Rhodes, M., Gelman, S. A., & Brickman, D. (2010). Children's attention to sample composition in learning, teaching and discovery. *Developmental Science, 13*(3), 421–429.
Rhodes, M., Leslie, S., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences, 109*(34), 13526–13531.
Rhodes, M., Leslie, S. J., Bianchi, L., & Chalik, L. (2018). The role of generic language in the early development of social categorization. *Child Development, 89*(1), 148–155.
Roberts, S. O., Ho, A. K., Rhodes, M., & Gelman, S. A. (2017). Making boundaries great again: Essentialism and support for boundary-enhancing initiatives. *Personality and Social Psychology Bulletin, 43*(12), 1643–1658.
Ronfard, S., Lane, J. D., Wang, M., & Harris, P. L. (2017). The impact of counter-perceptual testimony on children's categorization after a delay. *Journal of Experimental Child Psychology, 163*, 151–158.
Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development, 72*(4), 1054–1070.
Sabbagh, M. A., & Shafman, D. (2009). How children block learning from ignorant speakers. *Cognition, 112*(3), 415–422.
Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences, 16*(7), 382–389.
Schulz, L. E., Bonawitz, E. B., & Griffiths, T. (2007). Can being scared make your tummy ache? Naive theories, ambiguous evidence and preschoolers' causal inferences. *Developmental Psychology, 43*(5), 1124–1139.
Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development, 77*(2), 427–442.
Shipley, E. F., & Shepperson, B. (2006). Test sample selection by preschool children: Honoring diversity. *Memory & Cognition, 34*(7), 1444–1451.
Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental Science, 10*(3), 298–306.
Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review, 120*(4), 779–797.
Sobel, D. M., & Letourneau, S. M. (2018). Preschoolers' understanding of how others learn through action and instruction. *Child Development, 89*(3), 961–970.
Sobel, D. M., & Stricker, L. W. (2022). Messaging matters: Order of experience with messaging at a STEM-based museum exhibit influences children's engagement with challenging tasks. *Visitor Studies, 25*(1), 104–125.
Stock, H. R., Graham, S. A., & Chambers, C. G. (2009). Generic language and speaker confidence guide preschoolers' inferences about novel animate kinds. *Developmental Psychology, 45*, 884–888.
Taborda-Osorio, H., & Cheries, E. W. (2018). Infants' agent individuation: It's what's on the insides that counts. *Cognition, 175*, 11–19.
Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from mandarin speakers' early vocabularies. *Developmental Psychology, 32*(3), 492–504.
Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences, 104*(48), 19156–19159.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*(4), 629–640.
Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility. *Developmental Psychology, 47*(4), 1065–1077.
Tessler, M. H., Bridgers, S., & Tenenbaum, J. B. (2020). How many observations is one generic worth?. In *Proceedings of the Annual Conference of the Cognitive Science Society*.
Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*(6), 813–833.
Waxman, S. R., & Braun, I. (2005). Consistent (but not variable) names as invitations to form object categories: New evidence from 12-month-old infants. *Cognition, 95*(3), B59–B68.
Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive Psychology, 29*(3), 257–302.
Willard, A. K., Busch, J. T., Cullum, K. A., Letourneau, S. M., Sobel, D. M., Callanan, M., & Legare, C. H. (2019). Explain this, explore that: A study of parent–child interaction in a children's museum. *Child Development, 90*(5), e598–e617.
Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review, 126*(6), 841–864.
Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition, 112*(1), 97–104.
Zhao, H., Seibert, S. E., & Hills, G. E. (2005). The mediating role of self-efficacy in the development of entrepreneurial intentions. *Journal of Applied Psychology, 90*(6), 1265.