Learning from non-representative instances: Children's sample and population predictions

Jordan T. Thevenow-Harrison¹ and Charles W. Kalish

Department of Educational Psychology, University of Wisconsin–Madison

1025 W. Johnson Street

Madison, Wisconsin 53706-1796

Word count: 3,830

¹ Correspondence about this article should be addressed to Jordan T. Thevenow-Harrison, Email: jtth@education.wisc.edu.

Abstract

What do children learn from biased samples? Most samples people encounter are biased in some way, and responses to bias can distinguish among different theories of inductive inference. 67 four- to eight-year-old children learned to make conditional predictions about a set of sample items. They then made predictions about the properties of new instances or old instances from the training set. The experiment compared unbiased and biased sampling. Given unbiased samples, participants used what they learned to make predictions about population and sample instances. With biased samples, children were less accurate/confident about inferences about the population than the sample. Children used information in a biased sample to make predictions about items in that sample, but were less likely to generalize to new items than when samples were unbiased.

Keywords: Probability learning; Inductive inference; Cognitive development

Learning from non-representative instances: Children’s sample and population predictions

Learning from examples is both basic and central to forming knowledge of the world. There is significant debate about how best to characterize the psychological processes of learning from examples, and how to understand potential developmental changes in such processes (Colunga & Smith, 2008; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Xu, 2007). Some argue that learning is a process of forming associations (McClelland et al., 2010) while others argue that learning is best understood as evidential inference (Griffiths et al., 2010; Kalish & Thevenow-Harrison, 2014). Both positions are well characterized as “statistical learning.” Associative accounts focus on learning descriptive statistics: How do people notice patterns of co-occurrence in distributions of examples? Evidential accounts focus on learning inferential statistics: How do people use examples as evidence to draw conclusions about generative processes or populations (see Kalish, Kim, & Young, 2012)? Both associative and evidential accounts imply that people are sensitive to distributional features in samples. The accounts differ, however, in their treatment of the differences between samples and populations.

Consider the following situation. A fisherwoman catches a collection of fish that vary in size (big and small) and color (green and red). As she catches fish she notices all the big ones are red and all the small ones are green. How she notices and mentally represents the patterns (e.g., memory for specific examples, prototypes, conditional probabilities) are questions about the descriptive statistics she learns. Distributional features of the caught fish like the co-occurrences of sizes and colors, the number of examples, and the salience of the different attributes will affect the patterns she notices. Associative and evidential accounts agree here, but diverge in characterizing what happens when the fisher tries to use the patterns she noticed to make predictions. Prediction is a matching process in associative accounts: If she sees a big fish and wants to predict its color, she can match the target fish to the fish in her creel and return a judgment (e.g., it is more like the red fish than the green fish). In evidential accounts prediction is an inference to a generative

process: If she sees a big fish she considers what her experience has led her to conclude about the population of big fish. Critically, her conclusion may depend on two inferential features of the problem. The first inferential feature is the prediction target: Is she making a prediction about a fish from the creel, another from the pond she was fishing in, or about a fish from some other pond? The second feature is the representativeness of her sampling strategy: Was she using a particular kind of bait, or fishing from a location that might have biased her sample? Whether or how such inferential features affect predictions is where associative and evidential accounts diverge.

The specific evidential prediction is an interaction between prediction target and representativeness of sampling strategy. Strategy should matter for inference about new items from the population *but not* for inferences about old items from the sample. If the fisher is making a prediction about a fish from her creel, then the sampling strategy that produced the fish (the bait, the location, etc.) is irrelevant. One can guess that a big fish selected from the creel will be red without worrying about how that fish ended up in the creel in the first place. In contrast, predicting that a new big fish from the pond will be red *does* depend on the representativeness of the original sample. If the sampling strategy is biased then one cannot (easily) use the sample of caught fish to make a prediction about a new fish. This interactive prediction is the focus of the empirical study reported below: Will children use information about samples to make predictions about old (sample) examples but not new (population) examples when samples are non-representative, but use sample information for both old and new items when samples are representative?

Previous research suggests that young children are sensitive to both prediction target and sampling strategy. Several studies have found that young children and infants expect populations to resemble samples when the sampling strategy is representative (e.g., blindfolded selection) but not when it is not (e.g., deliberate selection; Xu & Tenenbaum, 2007; Xu & Denison, 2009). Young children are able to use a mismatch between sample and population to infer non-representative sampling (Kushnir, Xu, & Wellman, 2010).

Author and colleagues (2012) found that young school-aged children were better able to make predictions based on two distinct samples (e.g., fish from two creels) when the prediction targets came from the combined sample (from the creels) than when the targets were new instances (from the pond). Younger children did not distinguish between predictions about old versus new targets because they had difficulty integrating information from the two samples (i.e., tracking the descriptive statistics). The researchers interpreted this result as suggesting the presence of two samples with different distributions left children confused or suspicious of the sampling strategy.

In this experiment participants encounter examples with a correlation between two binary features. The process selecting the examples was either representative (e.g., random) or not (e.g., predetermined). Participants made a series of conditional predictions about individual examples either drawn from the sample or from the large population. Two groups of children participated in the study: four- to five-year-olds, and seven- to eight-year-olds. Author and colleagues (2012) provided evidence that young school-age children were sensitive to inferential relations of samples to population. Preschool-aged children did not display such sensitivity, perhaps because the frequency distributions involved in that prior study were too complex (the task involved combining two samples, one of which involved a probabilistic relation). One of the goals of the current study was to assess whether young children show sensitivity to inferential features when the frequency distributions are less complex (a perfect correlation).

We identify four hypotheses about learners' sensitivity to prediction target and sampling strategies when learning from samples. In the experiment reported below the prediction targets could be members of the sample (creel) or population (pond), and sampling strategy could be non-representative or representative. One hypothesis is to ignore both target and strategy and make predictions based on similarity alone. Another is to consider the nature of the target when computing similarity. An unknown fish from the creel is more like known fish from the creel than one from the pond. This strategy predicts

a main effect of target: the learner would make more accurate or confident predictions about sample than population examples. A third strategy focuses on quality of information. Perhaps learners ignore non-representative information: If the sample might be biased they stop learning, even about specific items in the sample. This strategy predicts more accurate and confident predictions given representative than non-representative sampling, a main effect of the sampling method. Finally the pattern consistent with evidential inference is to ignore sampling strategy when making predictions about the actual sample (fish from the creel) but favor representative samples when making predictions about the unseen population. In this case, participants will learn items in the sample given both sampling strategies, but will learn more about the population from representative samples.

Methods

Participants

The sample consisted of 35 four- to five-year-old (M age = 5.03, SD = 0.54) and 32 seven- to eight-year-old children (M age = 7.77, SD = 0.52) recruited from local schools in a mid-sized Midwestern college town. Younger and Older children were recruited the same way from the same population.

Design & Procedure

The experiment consisted of three phases: familiarization, training, and testing. In the familiarization phase the experimenter introduced children to the contents of two large canvas bags filled with toys one after the other in an order randomized between participants. While the child removed the toys from each bag the experimenter described the contents. While both bags had the same number of toys, one had many different kinds and colors of toys, and the other only two kinds and colors of toys. The familiarization phase introduced children to the idea that bags could contain different types and distributions of toys.

Figure 1. Illustration of training phase methods.

In the training phase, children saw items sampled by the experimenter from a third Population Bag, either at random (Random condition, $N_{\text{younger}} = 16$, $N_{\text{older}} = 16$) or from a printed list (Predetermined condition, $N_{\text{younger}} = 19$, $N_{\text{older}} = 16$). The method used to sample from the population varied between participants. Before the experimenter sampled an item from the bag he asked the participant to guess what the next toy would be. In the Predetermined condition the experimenter asked the participant to predict what came next on his list. The list consisted of color pictures of each toy in a line, and was wrapped in a piece of paper so each item could be revealed one at a time. In the Random condition the question concerned the next toy to be selected at random from the bag. Following the child's prediction, the experimenter selected a toy from the bag, either searching in the bag

to find the toy indicated on the list (Predetermined), or reaching in without looking (Random). The selected toy was set aside into a Sample Box in view of the participant throughout the training phase. In both conditions, items were actually selected in a pre-determined order (order was counterbalanced across participants). Because the list was seen to determine the kind and order of sampled items in the Predetermined condition, the training phase provides no information about the unobserved population of items left in the bag. Because the items seemed to have been sampled at random in the Random condition, the distribution of observed items does is informative about the remaining contents of the bag.

The sampling and prediction process repeated eight times. The toys sampled in both conditions are described in Table 1 below. For younger children toys had a perfect correlation between color and shape: Giraffes were blue, buses were green. Toys had a partial correlation for older children to make the task more challenging: Giraffes were green and blue, but buses were only green. Partial correlations between features were used with older children because pilot testing showed near-ceiling accuracy for perfect correlations, and older children are capable of using partial correlations to make inferences about populations (Thevenow-Harrison & Kalish, 2011).

Table 1

Count and distribution of features of toys seen by younger and older children.

Younger			Older	
Toy	Giraffe	Bus	Giraffe	Bus
Green	0	4	2	3
Blue	4	0	3	0

In the testing phase, children made a series of conditional predictions: They were given color and asked to predict shape (or vice versa). For both the Sample Box and the Population Bag in an order counterbalanced between participants, children were told they

were going to play a game about the toys in the Box and Bag and were asked to predict either color given shape or shape given color, then asked their confidence (*really sure*, *think so*, or *just guessing*). These questions were posed as hypotheticals, and children were not shown any stimuli while making these inferences (e.g., “If I were to reach into the box and pull out a giraffe, what color would it be?”“). To keep attention, after making their first four predictions (one for each feature) children were introduced to a cartoon child name Sally and asked four times what feature Sally should guess given the other feature. This testing phase was then repeated for the other toy source (Box or Bag). In total, children made 16 conditional predictions in which they were told one feature of a toy and asked to predict the other. Children made eight predictions about items in the Sample Box and eight about items in the Population Bag (counterbalanced order).

Measures

Participants provided two responses for each prediction: accuracy and confidence. We created a weighted “Inference Score,” consisting of accuracy (0 or 1) multiplied by confidence (1 – just guessing, 2 – think so, 3 – really sure). This weighs a child’s confident answers more than their unsure answers in a way raw accuracy does not and treats all incorrect answers the same regardless of confidence. Inference scores are a more sensitive measure of condition differences than accuracy alone. For example, a child might make the same predictions during Sample and Population tests but be more confident about the Sample. We also separately analyzed raw accuracy and confidence.

Results

Training

During training participants predicted which toy would be selected next. One measure of performance is the proportion of the last five training responses (out of eight) mentioning items the participant had seen up to that point. That is, were participants

limiting predictions to observed items? A Wilcoxon Mann-Whitney Rank Sum Test (Mann & Whitney, 1947) showed younger children in the Predetermined condition ($M = 3.26$, $SD = 1.56$) were no different than in the Random condition ($M = 2.38$, $SD = 1.99$, $Z = 0.91$, $p = .361$), and older children in the Predetermined condition ($M = 3.19$, $SD = 1.43$) were no different than in the Random condition ($M = 3.06$, $SD = 1.06$, $Z = 1.26$, $p = .206$). Children learned in both sampling conditions. Additionally, younger children’s last four predictions were more accurate than their first four, $M_{\text{first4}} = .09$, $SD_{\text{first4}} = .18$; $M_{\text{last4}} = .37$, $SD_{\text{last4}} = .27$, $t(34) = 5.16$, $p < .001$. The same was true for older children even with the partial correlation between color and shape, $M_{\text{first4}} = .05$, $SD_{\text{first4}} = .10$; $M_{\text{last4}} = .36$, $SD_{\text{last4}} = .25$, $t(31) = 6.39$, $p < .001$.

Testing

Younger children. An ANOVA with Inference Score (0 to 3) as the dependent variable and Target (Sample/Population) and Method (Random/Predetermined) as independent variables found no main effects or interactions that were significant. However, we found the hypothesized simple effect (see Figure 1), that Inference Scores would be higher for Sample than Population in the Predetermined condition, $F(1, 33) = 4.91$, $p = .034$, $\eta_p^2 = .129$, but not in the Random condition, $F(1, 33) < 1$, $p = .626$, $\eta_p^2 = .007$. Moreover Sample scores did not differ by Method condition, $F(1, 33) = 1.54$, $p = .223$, $\eta_p^2 = .045$, but Population scores were higher for children in the Random than in the Predetermined condition, $F(1, 33) = 4.74$, $p = .037$, $\eta_p^2 = .126$. Accuracy drove the difference between Sample and Population in the Predetermined condition (binomial GLMM, $z = 3.002$, $p = .003$). Also, confidence was generally higher in the random condition than the predetermined condition, $F(1, 33) = 9.54$, $p = .004$, $\eta_p^2 = .224$. No other main effects or contrasts were significant for accuracy or confidence.¹

Older children. In the analyses for older children we considered only predictions for

¹Contrasts were performed using the `lsmeans` R package on generalized linear mixed models from the `lme4` R package.

Figure 2. Inference scores during testing for younger and older children by condition, out of a maximum inference score of 3. For older children the figure shows only scores on perfectly correlated items. The simple effect of interest is illustrated by the differences between sample and population inferences in the predetermined condition for both age groups. Error bars represent one standard error.

the two perfect conditional probabilities present during training²: A bus will be green, and a blue thing will be a giraffe. Predictions when given giraffes and green things were uncertain and not included in these analyses. An ANOVA with the same structure as that for younger children revealed only an interaction between Target and Method, $F(1, 30) = 6.67$, $p = .015$, $\eta_p^2 = .182$. This effect was driven by our hypothesized simple effect: Sample

²For imperfect correlations, we calculated the proportion of guesses of the more frequent (3/5) feature. Children did not select the more frequent feature at rates significantly greater than chance for Sample or Population targets ($M_{\text{sample}} = .57$, $M_{\text{population}} = .56$) or given Random or Predetermined select ($M_{\text{random}} = .61$, $M_{\text{predetermined}} = .52$, all n.s., 2-tailed t-tests).

inference scores were higher than Population scores, $F(1, 30) = 8.88$, $p = .006$, $\eta_p^2 = .228$ in the Predetermined condition. Sample and Population scores did not differ in the Random condition, $F(1, 30) < 1$, $p = .506$, $\eta_p^2 = .015$. In contrast to younger children, older children's Population scores did not differ by Target condition, $F(1, 30) < 1$. Sample inferences were higher in the Predetermined condition, $F(1, 30) = 4.14$, $p = .05$, $\eta_p^2 = .121$. One interpretation of this effect is that older children were more attentive to stimuli in the Predetermined condition. Older children's inference scores were driven by confidence. In the Predetermined condition judgements on the Sample were higher than the Population, $F(1, 30) = 7.45$, $p = .011$, $\eta_p^2 = .199$. There were no accuracy differences in pairwise contrasts.

Summary

Younger and older children both made weaker inferences to new examples from a population than to old examples from the sample, but only when the sampling process was non-representative. This interaction effect is consistent with the evidential hypothesis. However, the age groups instantiated the hypothesized effect in different ways. For younger children, Predetermined-Population inference scores were lower than all other scores. These participants generalized, but not for this critical cell. For older children there seemed to be an effect whereby scores in the predetermined condition were higher. Older children were better in the predetermined condition overall, which is not predicted by any of the four alternatives.

Discussion

Children made weaker or less accurate predictions to new population examples than to old sample examples only when the sampling strategy that produced the old examples was non-representative. When the sampling process was representative, children were equally confident/accurate for both old and new examples. The pattern for younger children was clear: Sampling strategy only made a difference for predictions to new items.

When making predictions about old members of the sample, the process that generated the sample was irrelevant. If children are treating examples as samples to draw inferences about a population from, then we would expect them to be reluctant to generalize to that population when the sampling method is non-representative. Representativeness should not affect predictions about sample examples. This hypothesis was contrasted with two hypotheses about independent main effects: A strategy where children may make weaker predictions about population items than samples, or where a child may make weaker predictions given non-representative samples than representative ones. There were no main effects of either target, predicted by the former, or sampling method, predicted by the latter.

One conclusion is that young children do treat predictions about new items as different from predictions about old items. It is tempting to think of judgments made about the old (sample) items as mere recall, but recall of experienced patterns is central to both old and new predictions (Kalish & Thevenow-Harrison, 2014). The question addressed in the current study is when predictions go beyond reliance on memory for patterns observed in the past. When the to-be-predicted item is old, then answering based on past patterns is always appropriate. When the to-be-predicted item is new, then the relevance of those previously experienced patterns depends on the representativeness of the earlier samples. This sensitivity to representativeness is just what was observed in the current experiment, though older and younger children were sensitive to representativeness in different ways.

Older children's pattern of results support our hypothesis that there would be a difference between sample and population inferences in the predetermined but not the random condition, but not in the way we expected. One explanation is that the older children were just paying more attention in the predetermined condition. For example, older children's patterns of responses look very similar to younger children's but with a constant added in the predetermined. Older children may have treated the predetermined list as a pedagogical cue. Not all non-random samples are untrustworthy. In a pedagogical

context an efficient teacher selects examples that reflect certain properties that are important about the population those examples belong to (referred to as pedagogical sampling, see Shafto & Goodman, 2008). The hypothesis is that older children placed more importance on intentional (Predetermined) than on random samples. It is important to note that older children were not simply preferring pedagogical/predetermined samples. These children were more likely to use such samples as the basis of predictions for old (sample) items than for new (population) items.

Our experiment and many others suggest that children interpret examples as samples that generalize to populations. If so, then children face the problem of identifying just which population is represented in the samples they encounter. The current experiment contrasted blind selection with a predetermined list. This contrast between sampling strategies was effective, but remains open to multiple interpretations (e.g., was one strategy pedagogical). Clearly there are many ways to generate representative and non-representative samples (e.g., mechanical bias, deception, see Thevenow-Harrison & Kalish, 2011). Future research should explore just how children recognize representative and non-representative samples and, critically, how such differences affect the inferences made. As Fiedler (2008) notes, no one sampling strategy is optimal for all inferential tasks.

Previous work has shown that infants and young children are sensitive to bias in sample selection; they distinguish between representative and non-representative samples (Xu & Denison, 2009). In our design, this prior finding would be a main effect of sampling method. The point of the current study, however, is that children did not always make stronger inferences with representative than non-representative samples. Sampling method only mattered when target items were novel, not part of the original sample. In most prior research it is unclear (to the participant) whether the items presented at test are members of the training set or not (see Kalish et al., 2012, for discussion). If children’s sensitivity to sampling method reflects something like assessment of evidence, then we would expect method to matter for new but not old items. In contrast, if sensitivity to sampling reflects

some more general bias or dispreference (e.g., against tricky or non-random selections) then we would expect the same results for both old and new items. Our results suggest that children are considering the evidential value of information they acquire.

These experiments lend support to the idea that children consider the representativeness of a sample and use that information when generalizing to a new instance drawn from the same population. We asked under what conditions do children consider evidence when making an inference. Consistent with other work we found that children pay attention to inferential features, such as sampling strategy and inference target. However, we also found they do not *always* attend to inferential features. Both associative and evidential accounts of children's inference learning can accommodate that children consider whether a stimulus old or new, and whether or not a sample is representative. The critical point is that these two features are combined in a rational way: the novelty of a stimulus only matters when the method used to sample it is not representative. What will move the field forward is not *whether* children never or always attend to the inferential features of a situation, but rather *when* they do so.

References

- Colunga, E. & Smith, L. B. (2008). Knowledge embedded in process: the self-organization of skilled noun learning. *Developmental Science*, 11(2), 195–203.
- Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 186–203.
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357–364.
- Kalish, C. W., Kim, S., & Young, A. G. (2012). How young children learn from examples: descriptive and inferential problems. *Cognitive Science*, 36(8), 1427–1448.
- Kalish, C. W. & Thevenow-Harrison, J. T. (2014). Descriptive and inferential problems of induction: toward a common framework. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 1–39). Elsevier.
- Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, 21(8), 1134–1140.
- Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1), 50–60.
- McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8), 348–356.
- Shafto, P. & Goodman, N. (2008). Teaching games: statistical sampling assumptions for learning in pedagogical situations. *Proceedings of the thirtieth annual conference of the cognitive science society*, 1632–1637.

- Thevenow-Harrison, J. T. & Kalish, C. W. (2011). What do children learn through constrained sampling? Sample but not population statistics. Poster presented at the annual meeting of the Cognitive Science Society, Boston, MA.
- Xu, F. (2007). Rational statistical inference and cognitive development. *The innate mind: Foundations and the future*, 3, 199–215.
- Xu, F. & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, 112(1), 97–104.
- Xu, F. & Tenenbaum, J. (2007). Sensitivity to sampling in bayesian word learning. *Developmental Science*, 10(3), 288–297.