The Impact of Statistical Training on Children's Inductive Reasoning

Susan L. Stanley (shlocke@uwm.edu)
Chris A. Lawson (lawson2@uwm.edu)
University of Wisconsin – Milwaukee
Department of Educational Psychology
Milwaukee, WI USA

Abstract

Regardless of age there are mixed findings concerning the extent to which individuals utilize statistical features of input to make inductive inferences. Direct instruction seems to be one important factor in linking one's understanding of statistical properties with their reasoning. In the present study we examined the extent to which explicit training on some statistical principles would influence preschoolers' inductive reasoning. The results indicate that a short training about random selection and the match between samples and populations increased children's use of these principles to make inductive generalizations. Critically, the training effects were observed in a different domain than was presented in the training and for statistical principles not presented in the training. Thus, the present results suggest that the training had a broad impact on children's reasoning. These results have important implications for understanding the nature of the statistical principles employed during induction.

Keywords: statistical training; sample size; inductive reasoning; cognitive development; generalization

Introduction

Most decisions involve inductive generalizations, the use of specific information to draw general conclusions. Induction is interesting to cognitive scientists because, despite the apparent uncertainty of predictions, people are quite good inductivists. Among other things, our inductive reasoning is anchored by careful attention to statistical features of evidence. For example, recent advances in Bayesian models of induction suggest the inductive inferences of even the youngest learners are closely aligned with the most (statistically) optimal predictions (e.g., Xu & Kushnir, 2013). However, there are many cases in which children (Lopez, Gelman, Gutheil, & Smith, 1992) and adults (Kahneman, Slovis, & Tversky, 1982) fail to adhere to statistical properties of evidence and instead rely on other sources of information to make predictions. The present study focused on conditions in which children have yet to acquire the statistical principle that will support induction. We were particularly interested in understanding whether children could acquire such skills with minimal instruction, and whether learning about specific statistical principles (e.g., random sampling) would have a narrow effect on induction (e.g., to inferences regarding random selection) or a broader effect (e.g., to inferences concerning other statistical principles such as sample size).

There is mixed evidence regarding the extent to which young children incorporate statistical evidence into their inductive inferences. On the one hand, recent studies indicate that prior to their first birthday infants exhibit an intuitive set of skills that enable them to generate expectations in-line with the statistical features of evidence (Xu & Garcia, 2008; Teglas, Girotto, Gonzalez, & Bonatti, 2007). For example, 8-month-olds expect a randomly selected sample will have the same distribution of white and red balls as the population from which it was chosen.

In addition to exhibiting the intuition that samples match the distributions of the populations from which they were chosen, young infants also appear to be aware of the likely outcomes of the procedures used to yield samples. For example, infants expect the distribution of a sample to be similar to that of the population from which it was drawn when an actor randomly selected items from the sample, but not when the actor engaged in deliberate sampling (Xu & Tenenbaum, 2007; Denison, Reed, & Xu, 2011; Gweon & Schulz, 2011). Overall, by the end of the first year infants appear to exhibit some powerful expectations about the statistical properties of evidence.

However, children appear limited in their ability to effectively use statistical properties to evaluate which samples provide the best support for induction. Some studies have shown that children are insensitive to the sample size principle of induction until after age 8 or 9 years of age (Gutheil & Gelman, 1997; Li, Cao, Li, Li, & Deak, 2009; Lopez et al., 1992; cf. Lawson, 2014; Lawson & Fisher, 2011). For example, Gutheil and Gelman (1997) presented children evidence about a single exemplar (e.g., "This butterfly has blue eyes") and a sample of five exemplars (e.g., "These five butterflies have gray eyes.") and asked them to generalize a property from one of the samples to an evidence target (e.g., "Do you think this butterfly has blue eyes like this butterfly, or gray eyes like these butterflies?"). Children younger than 8 years of age responded randomly, indicating they failed to recognize that the larger sample provided better evidence to support induction. In contrast, children older than 8 consistently prefer to generalize from large, rather than small samples of evidence (see also Li et al., 2009; Lopez et al., 1992).

These mixed findings do not necessarily reflect developmental change in inductive skills. Indeed, there is considerable variability in adults' use of sample size to make inductive generalization (e.g., Sedlmeier & Gigerenzer, 1997). Classic studies in the heuristics and biases literature suggest that adults fail to obey the sample size principle of induction when reasoning about everyday problems or when evidence is framed in probabilistic terms (e.g., Kahneman & Tversky, 1972). However, when problems are posed in a format that makes sample size information salient to them (e.g., frequencies) adults recognize the inductive value of larger samples (Sedlemeier & Gigerenzer, 1997). Likewise, adults obey the sample size principle to generalize properties in a category-based induction task (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990).

One conclusion from this work is that the ability to use statistical features in the evidence depends, at least in part, on an understanding of the inductive value of these features. For example, older children and adults may understand the value of larger samples for making predictions, but fail to see the connection to making judgments about variability. Similarly, because infant studies examine preferential looking, it is not entirely clear how, or whether, early emerging expectations about statistical principles impacts performance on evaluative reasoning tasks.

From this perspective it is natural to ask how people develop an understanding of the inductive value of statistical evidence. One answer is that such an understanding is the product of direct instruction. In an extensive body of work, Nisbett and colleagues showed that training adults to attend to statistical features they otherwise ignore caused adults to incorporate these features into their inductive inferences (Nisbett, Krantz, Jepson, & Kunda, 1983; Fong, Krantz, & Nisbett, 1986; Nisbett, Fong, Lehman, & Cheng, 1987; Lehman & Nisbett, 1990; Fong & Nisbett, 1991). For example, Fong et al. (1986) provided participants in the training condition with a four-page description of the law of large numbers and the statistical principle of sampling. The written descriptions of the concepts were supplemented with a live demonstration of the law of large numbers using blue and red gumballs in a glass container. Samples of 1, 4, and 25 gumballs were drawn from the container and results were recorded on a blackboard. The experimenter noted sample size differences by describing that the larger sample deviates less from the population of gumballs than did the smaller sample. Those who received the training were more likely to use statistical reasoning when asked to solve problems that covered a wide range of domains, such as decisions made at one's job, buying a car, and choosing which college to attend. Thus, statistical training seems to have general, rather than specific, effects on reasoning.

Only one study examined the effects of statistical training on children's reasoning. Kosonen and Winne (1995) assessed college, high school, and middle school students' skills in solving everyday problems after being taught the law of large numbers in a regular classroom setting. After reading a text that introduced the concepts of the law of large numbers, students were shown a live demonstration of the outcomes of random sampling. The demonstrator drew gumballs from a large urn and recorded the colors on a blackboard. After the training, students at all grade levels demonstrated improvements in their use of statistical reasoning to solve real-life scenarios. Participants were able to transfer the statistical rules they were taught to a broad range of topics such as decisions of hiring prospective employees, choosing a restaurant, playing a board game, and judging someone's personality after a quick interaction.

The present study extends this work by examining the effects of statistical training on preschoolers' use of statistical information to make inductive inferences. We expected that providing young children with a short lesson about some basic statistical principles (e.g., random sampling) would increase their use of statistical properties of evidence to make inductive generalizations. In particular, we were interested to see whether teaching children about the likely outcomes of random sampling and the distributional likeness between samples and populations would cause them to use these principles when making inductive generalizations.

An additional goal of this study was to examine the scope of the impact of statistical training on children's reasoning. In our study we used a container of ping pong balls to teach children statistical principles. Does a lesson of this sort increase children's attention to other statistical properties, such as sample size, to make inferences? In addition to including items that required attention to sample size, the test items examined generality of training effects by asking children to make predictions in a range of domains involving social actors and biological categories. Does learning about abstract statistical principles applied to a sample of ping pong balls generalize to reasoning about everyday problems that involve people and animals?

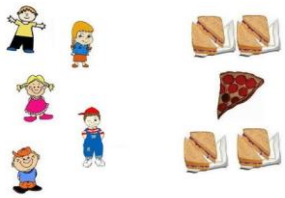
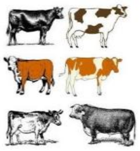
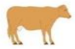
Experiment

Method

Participants. Fifteen preschoolers (M= 5.01, SD = 0.54; 8 females, 7 males) participated in the training group and fourteen preschoolers (*M*=5.23, *SD*=.46; 8 females, 6 males) participated in the control group. Participants were recruited from local preschools. They were from diverse racial backgrounds and were representative of the city of Milwaukee. Schools received a small monetary gift for their participation.

Design and Materials. The training condition employed a pre-posttest design that was conducted over the course of three days. Over the three days, participants responded to three types of questions that involved an appreciation of the following statistical properties of evidence. These questions were modeled after those used in the adult and

Table 1
Experimental design.

Item Type	Exemplars	Script
Random Sampling		There are five peanut butter cookies and two chocolate chip cookies in this box. Donna reached into the box without looking and picked the first cookie she touched. Which cookie do you think Donna picked – peanut butter or chocolate chip?
Sample-to-Population		Ms. Hansen was trying to decide what to get for lunch for the whole class. When she asked the first five students who came to school four students said they wanted peanut butter and jelly sandwiches and one student said they wanted pizza. What about all the other students? Will more students want peanut butter and jelly or will more students want pizza for lunch?
Sample Size		These cows have Type A blood.  This cow has Type B blood. Do you think this cow has Type A or Type B blood?

Note. Five questions of each item type were presented in random order to participants at the pretest and posttest.

developmental literature on statistical reasoning. The following types of items were administered to children (see Table 1 for sample items):

1. *Random sampling* items measured the extent to which children expected that random selection of an item would yield a high probability instance.

2. *Sample-to-population* items measured whether children recognized that the distribution of a sample would match that of the population from which it was drawn.

3. *Sample size* items assessed whether children recognized that a large sample of evidence provided better support for induction than a small sample of evidence.

Procedure. All testing sessions were conducted in a quiet location at each child's preschool. In the Training group the experiment was conducted over three days, each of which is described below.

Day 1 – Pretest: Children were given five questions from each of the three item types (random sample, sample-to-population, and sample size) yielding a total of 15 pretest questions. Participants were told that there were no wrong answers. Each item was accompanied by a picture to help children pay attention during the task.

Day 2 - Training: During the training session a box containing white and orange ping pong balls was used to demonstrate the consequences of random sampling and the similarities between a sample and the population from which it was selected. The population included 140 white ping pong balls and 12 orange ping pong balls all of which were presented in a large box with a clear window so participants could see the ping pong balls. A smaller, clear box was used to hold the sample of ping pong balls that was drawn from the larger box.

During the training, the researcher reached into the large box without looking and picked out a ping pong ball. The researcher repeated this procedure five times and placed each ping pong ball in the smaller box. The large box was rigged with a small compartment (not observable to the participants) to assure the experimenter could select one orange ball and four white balls from the box. After the selection of each ping pong ball from the larger sample, the researcher mentioned the similarity between the small box and large box, describing that there were mostly white balls in both of the boxes. It was also highlighted that the researcher did not look while reaching into the large box, but instead, "just reached in and grabbed the first ball I touched." This procedure was repeated until there was a sample of five ping pong balls (four white and one orange). After the five balls had been selected the researcher highlighted that the sample and the population both looked similar because they both had more white balls than orange balls. The researcher than noted that if she wanted to she could have looked inside the box and chosen orange balls, but that by just picking the balls "without looking" both boxes had more white balls than orange halls.

After the training, participants were asked two random sample questions and two sample-to-population questions.

After they gave their response, the experimenter provided feedback, drawing the children's attention back to the ping pong balls to help identify the correct response and to mention the principle that justifies this response. The training session took between 10-15 minutes for each participant.

Day 3 - Posttest: Similar to Day 1 except participants were given a second set of test items from each of the three question types (five from each) presented in random order.

The Control condition used the same design and materials as the Training condition with the exception that participants in this group did not receive any training or feedback on responses to questions.

In the training condition all three sessions were conducted within 5 days; critically, there were no more than 3 days between the training and posttest sessions. In the control condition the testing took place over two consecutive days instead of three.

Two sets of items were designed for presentation in the pretest and posttest. Each set contained a randomly selected set of questions from the three item types. The sets were counterbalanced so that an approximately equal number of sets appeared as the pretest of posttest. Initial analyses confirmed that neither set of questions was more prone to elicit adherence to the statistical principles in the evidence.

Results

Responses were scored on the basis of whether participants obeyed the statistical principles in each item. A "1" was given for a response that was consistent with the statistical principle (e.g., preference to generalize from the large, rather than small sample; judgment that the population would yield a sample that resembled the population; and responding that random sampling would yield a high probability outcome), and a "0" was given for a response that was inconsistent with the statistical principle. To test the two main predictions the mean statistics-based responses were submitted to a Day (Pretest, Posttest) x Item Type (sample-to-population, random sampling, sample size) analysis of variance with repeated measures. Consistent with the hypothesis that training would increase children's statistics-based responses prediction there was a main effect of Day, $F(1, 27) = 12.12$, $p = 0.002$, $\eta^2 = 0.31$ and a Day by Condition interaction, $F(1, 27) = 11.29$, $p = 0.002$, $\eta^2 = 0.26$ (see Figure 1). This interaction was due to an increased rate of statistics-based responses during the posttest in the Training group than in the Control group. This result indicates that the training, rather than extraneous factors associated with the prolonged testing, significantly increased children's use of statistical principles to make inductive inferences.

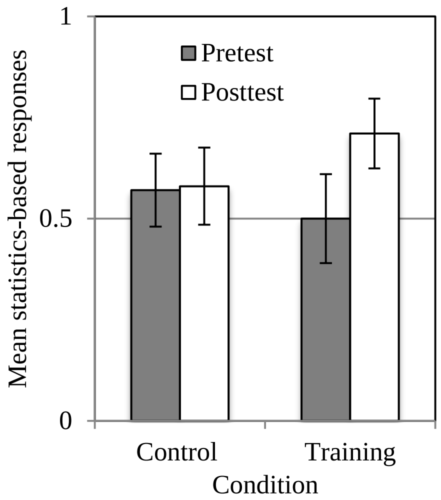


Fig.1. Mean statistics-based responses during Pretest and Posttest in the Control and Training conditions. Bars indicate one standard error from the mean.

The second hypothesis was that the training effects would generalize beyond the principles introduced during training. Some support for this prediction comes from the overall ANOVA for which the Item by Condition interaction was not significant, $F < 1.0$, $\eta^2 < 0.10$. Thus, the heightened rate of statistics-based responses in the posttest was not due to the two items that included statistical principles embedded in the training (random selection, sample-to-population), but instead was consistent across all items.

Table 2
Mean statistics-based responses for each of the item types in the Training and Control conditions

	Pretest		Posttest	
Item Type	*M*	*SD*	*M*	*SD*
		Training (n =15)		
Random Sample	0.45	0.20	0.75	0.26
Sample-to-Population	0.56	0.23	0.71	0.28
Sample Size	0.49	0.26	0.65	0.21
		Control (n = 14)		
Random Sample	0.57	0.31	0.56	0.28
Sample-to-Population	0.64	0.26	0.64	0.21
Sample Size	0.50	0.29	0.53	0.24

A more direct test of this second hypothesis involved one-way ANOVAs to determine if there was a significant difference from pretest to posttest for each of the items (Table 2). The tests revealed no significant difference among the item types in the Control group, all pairs revealed *F(1, 27)* < 0.061, *p* > 0.80. Among the item types in the Training group, random sampling, *F*(1, 28) = 11.93, *p* = 0.002, showed a significant difference from pretest to posttest. However, the sample-to-population, *F*(1,28) = 2.448, *p*=0.129, and sample size, *F*(1, 28) = 3.476, *p* = 0.073, did not reveal any significant differences over time.

Finally, we looked at comparisons to chance (*M*=.50) to examine the consistency of responses before and after the training. Results of one-sample *t*-tests showed responses to all item types were consistently better than chance after training (random sampling, *t*(14) = 3.73, *p* = 0.002; sample-to-population, *t*(14) = 2.84, *p* = 0.013; sample size, *t*(14) = 2.88, *p* = 0.012). The participants in the Control group consistently answered better than chance on the posttest (*t*(13) = 2.543, *p* = 0.025) on the sample-to-population items. The results on the pretest (*t*(13) = 2.04, *p* = 0.062) for sample-to-population items were trending toward significance. However, the responses on the random sample and sample size items were no different from chance on the pre- and posttest in the Control group.

Discussion

There are mixed reports on the extent to which children and adults are able to incorporate statistical properties into their inductive decisions. One explanation for these mixed findings is that they reflect differences in exposure to the rules that govern the use of statistical information. Prior work by Nisbett and colleagues supports this interpretation – the extent to which adults incorporate statistical principles into their inductive decisions is mediated by training on these statistical principles (Fong et al., 1986; Nisbett et al., 1983; 1987). The goal of the present study was to examine whether similar effects could be demonstrated in preschoolers.

The results from this study confirmed that a brief training on the statistical principles of random sampling and the match between samples and the populations from which they were selected influenced children's use of these principles to make inductive judgments. The observed training effects are particularly interesting for two reasons. First, because the content of the training (e.g., lottery type events with ping pong balls) and the test items (e.g., outcomes in social and biological scenarios) were quite different, these training effects suggest there was considerable transfer of the statistical principle from training to posttest.

The second, related, observation is that the training effects had a broad effect on induction. Thus, in addition to enhancing performance on items that required awareness of sample-to-population statistics and random selection the training influenced performance on items that required use of the sample size principle. One interpretation of this finding is that some aspect of the training primed children's attention to sample size. For example, comparison of the sample and population during training may have highlighted sample size differences. Another interpretation is that rather than teaching children about specific properties of evidence, the training taught children about general principles of reasoning. That is, the training may have forced children to focus on the composition of the samples and methods for accessing and using evidence. Though there is little doubt the training influenced children's inductive judgments, important questions remain concerning why the training had these effects.

Another finding in this study is that children in the Control group consistently answered better than chance on the sample-to-population items. These results go along with previous studies that found that young infants anticipate a sample will have a similar distribution as the population (Denison et al., 2011; Teglas et al., 2007). The novel finding here is that the ability to detect the sample most likely to be drawn from a population appears to be an important insight for drawing inductive decisions. However, it is important to note that while children consistently used the sample-to-population rule in the Control condition, they only did so after the training in the other group. Thus, more work is needed to understand the extent to which children rely on the match between sample and populations as a basis for inductive generalization.

Indeed, rather than contesting the view that young learners are intuitive statisticians (e.g., Teglas et al., 2007; Xu & Garcia, 2008) the present study challenges the meaningfulness of such a designation. On the one hand the intuitive statistics perspective has been criticized on the basis that the methods used to measure infants' recognition of some inductive principles (e.g., matching samples to populations) actually assess perceptual skills (Lawson & Rakison, 2013). Moreover, because preschoolers in the present study failed to respond in-line with any of the statistical principles during pretest it remains to be seen just how well they are able to incorporate statistical properties of evidence into their inductive decisions. Certainly the inability to incorporate statistical properties into their inferences does not mean children are unable to recognize these features in the evidence. Instead, the point is that there is a considerable gap in the literature on statistical reasoning in young children, such that the characterization of infants as gifted statisticians (e.g., Xu & Garcia, 2008) needs to be reconciled with other work showing the limitations in the statistical reasoning of young children (e.g., Gutheil & Gelman, 1997) and adults (Kahneman et al., 1982; Sedlmeier & Gigernezr, 1997).

The present study might also have practical implications for education. National standards dictate that students receive direct instruction about statistics starting by grade six (National Governors Association Center for Best Practices, 2010). One explanation for why this training comes so late is that statistical principles are beyond the grasp of younger children. However, such a conclusion seems unwarranted. In fact, one interpretation of the present results is that young children can benefit from early statistics instruction. One question for future research is to examine the overall effects of such early instruction. For example, does instruction about statistical properties of evidence have an influence on performance on mathematics or an overall impact on critical thinking skills? And does this training have lasting effects that extend beyond the few days as was observed in the present studies? Many topics in the mathematics curriculum are necessary for mastery; however they are not applicable to most people's daily lives. Probability and statistics are constantly being used every day. Statistics deals with the risks, rewards, and randomness of situations that people may encounter. Teaching children about probability and statistics should be at the forefront of the mathematics curriculum since it is something they can apply to their day-to-day lives.

Acknowledgments

This study would not be possible without the cooperation of the teachers, parents, and children of the Ebenezer Child Care centers in Wisconsin. Thanks also to Chelsea Kieler and Lily Cabrera for assistance with data collection.

References

Denison, S., Reed, C., & Xu, F. (2011). The emergence of probabilistic reasoning in very young infants. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society,* Austin, TX: Cognitive Science Society. doi: 10.1037/a0028278.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18, 253 – 292.*

Fong, G. T., Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120, 34-45.*

Gutheil, G. & Gelman, S. A. (1997). Children's use of sample size and diversity information within basic-level categories. *Journal of Experimental Child Psychology, 64, 159 – 174.*

Gweon, H. & Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, 332, 1524. doi: 10.1126/science.1204493

Jacobs, J. E. & Narloch, R. H. (2001). Children's use of sample size and variability to make social inferences. *Applied Developmental Psychology, 22, 311 – 331.*

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3, 430 – 454.* doi:10.1016/0010-0285(72)90016-3

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and Biases. Cambridge University Press.

Kosonen, P. & Winne, P.H. (1995). Effects of teaching statistical laws on reasoning about everyday problems. *Journal of Education Psychology, 87, 33 – 46.* doi: 10.1037/0022-0663.87.1.33

Lawson, C.A. (2014). Three-year-olds obey the sample size principle of induction: The influence of evidence presentation and sample size disparity on young children's generalizations. *Journal of Experimental Child Psychology*.

Lawson, C. A. & Fisher, A. V. (2011). It's in the sample: The effects of sample size and sample diversity on the breadth of inductive generalization. *Journal of Experimental Child Psychology, 110, 499 – 519.*

Lawson, C.A., & Rakison, D.H. (2013). Expectations about single event probabilities in the first year of life: The influence of perceptual and statistical information. *Infancy, 18*, 961–982. doi: 10.1111/infa.12014

Li, F., Cao, B., Li, Y., Li, H., & Deák, G. (2009). The law of large numbers in children's diversity-based reasoning. *Thinking and Reasoning*, 15, 388–404.

Lopez, A., Gelman, S.A., Gutheil, G., & Smith, E.E. (1992). The development of category-based induction. *Child Development, 63, 1070 – 1090.* doi: 10.1111/j.1467-8624.1992.tb01681.x

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). Common Core State Standards. Washington D.C.

Nisbett, R. E., Fong, G. T., Lehman, D. R., Cheng, P. W. (1987). Teaching reasoning. *Science, 238, 625 – 631.* doi: 10.3758/BF03204711

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90, 339 – 363.*

Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185-200.

Sedlmeier, P. & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10, 33-51.*

Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. (2007). Intuitions of probabilities shape expectations the future at 12 months and beyond. *Proceedings of the National Academy of Sciences of the United States of America, 104, 19156-19159.* doi: 10.1073/pnas.0700271104

Xu, F., & Kushnir, T. (2013). Infants are rational constructive learners. *Current Directions in Psychological Science, 22,* 28-32.

Xu, F. & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 5012-5015. doi: 10.1073/pnas.0704450105

Xu, F. and Tenenbaum, J. B. (2007), Sensitivity to sampling in Bayesian word learning. *Developmental Science, 10*, 288–297. DOI: 10.1111/j.1467-7687.2007.00590.x