Reasoning from Data: The Effect of Sample Size and Variability on Children's and Adults' Conclusions

Amy M. Masnick (masnick@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh PA 15213

Bradley J. Morris (bjmorris@pitt.edu)
Learning, Research, & Development Center, University of Pittsburgh
Pittsburgh PA 15260

Abstract

Interpretation of data is a critical part of scientific experimentation because it involves applying one's background theoretical knowledge to the characteristics of the data. Though many researchers have examined the impact of background knowledge, few have considered the impact of the characteristics of the data in making decisions. In this study, we presented 3rd graders, 6th graders, and college undergraduates with a series of datasets that varied in sample size, consistency in data pairs and variability relative to the mean. We found that at all ages, participants showed sensitivity to sample size and whether or not there were overlapping data points in comparative datasets, but that there were age differences in the justifications used and in conclusions drawn from the data.

Interpretation of data is a critical part of scientific experimentation. Expectations about features of the data have been suggested as an important component in assessing data (Kahneman & Tversky, 1973). These expectations are based both on theoretical knowledge about the domain under consideration and on features of the data itself. While a large body of research in scientific thinking examines the influence of domain theory on the evaluation of data (e.g., Klahr, 2000; Koslowski, 1996; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995), little is known about how the characteristics of data influence how children and adults interpret it.

An important component of science is distinguishing real effects from error, or effects caused by factors other than the ones being explored. In the science laboratory, statistics is a vital tool to help make these decisions. When there are differences that are highly unlikely to occur by chance, scientists can feel more confident about drawing conclusions from data.

In daily life, we regularly make decisions about evidence without the aid of formal statistics. In such cases, we resort to relying on theory and expectations. However, there are many situations in which we do not have strong background information, and thus only have evidence based in the data. Elementary school students seem likely to have an especially large handicap in evaluating data – they have a smaller knowledge base about the world and also have less formal knowledge about statistics and its applications.

Students in elementary school are beginning to learn about experimentation and data interpretation, and third through sixth grade is a time of important increases in understanding of basic science fundamentals, such as the control of variables strategy (e.g., Chen & Klahr, 1999). In addition, elementary school teachers routinely assign children to perform repeated trials of events, explaining that this is how science is done (Klahr, Chen & Toth, 2001). In evaluating data in and out of the classroom when children do not know formal statistical techniques, we expect them to rely on their informal knowledge of the area.

But what constitutes "informal" notions of statistical reasoning? We suggest two components: expectations about data distribution and expectations about the influence of sample size. Some research that has examined expectations for the distribution of data has looked at probability estimates. For example, when given data about a series of coin flips, participants expected that a coin would land on "heads" every other flip (Gilovich, 1991). This suggests that the participants had an implicit expectation of the distribution of data in a series of coin flips and that the judgment of "randomness" was (at least in part) based on a mapping between expectations and data patterns. More recently, some have argued that children as young as five or six have a functional understanding of probability (Schlottman, 2001).

Although there is related research in several areas, few studies focus explicitly on the characteristics of the data and the effects this focus has on conclusions. There is some evidence that children at different ages do recognize different properties of datasets, and that this recognition in turn affects the conclusions they draw. For example, Jacobs and Narloch (2001) found that children as young as seven could use sample size and variability information in inferring the likely frequency of a future event. The differences in variability were based on prior knowledge of base rates (i.e., how many elephants have two eyes, compared to how many birds are a specific color). The sample sizes used in this study varied dramatically, with either 1, 3, or 30 instances of an event before the

participant was asked to infer likelihood of other instances of the event, so it is unclear what sample size leads children to feel confident in their predictions.

However, there is also some evidence that children at ages 11 are still struggling to understand the value of repeated measurements within the context of a school science laboratory (Lubben & Millar, 1996). Some children at this age believe repeated measurements are important, but 18% thought that repeated measures are useful because they accommodate scatter in the data.

There is also some evidence that children can distinguish different kinds of variability. Masnick and Klahr (2001) examined second and fourth graders performing experiments in which two balls were simultaneously rolled down ramps and the distance each travels was measured. The children expected that on a new trial using the same experimental set-up, the relative positions of the two balls would remain the same, but that the precise location of each ball might be different. That is, they were able to make a distinction between small differences in individual data points and larger differences in sample means.

Students' expectations about the essential features of data and the features of a specific dataset may allow them to recognize data as consistent or inconsistent with their expectations about its distributions. These expectations may in turn guide decisions about the usefulness of the data and the extent to which the data are relevant to explanatory theories. Thus, the characteristics of the data partly determine the extent to which they are used to guide the formation or modification of explanatory theories.

One related body of research has examined how children use data (in this case covariation between events) to detect causal relationships between elements (Shaklee & Mims, 1981; Kuhn, 1989). These studies looked at how children evaluate evidence when events occur together all the time, some of the time, or none of the time.

In a series of studies by Shaklee and her colleagues, students in grades 2-8 and adults were presented with data about two events (e.g., plant growth-healthy/unhealthy and use of bug spray- yes/no) in a 2x2 contingency table. From these data, students were asked to determine the nature of the causal relationship between the events (i.e., presence and direction of relationship). A majority of participants at all ages did not use conditional probability rules to determine covariation, yet many children used a strategy for summing the diagonals in the contingency table (Shaklee & Paszek, 1985). However both children and adults could use a conditional probability rule if instructed (Shaklee, Holt, Elek, & Hall, 1988).

Kuhn and her colleagues (Kuhn, Amsel, & O'Loughlin, 1988; Kuhn et al., 1995) extended this line of research and examined the effect of detecting covariation on the participant's prior beliefs about an event. For example Kuhn et al. (1988) interviewed sixth and ninth grade children about the relationship between consuming different types of foods and catching colds. Information about each child's prior beliefs was used to provide each child with two sets of data: one that confirmed their prior beliefs and one that disconfirmed their prior beliefs. The researchers argued that children did not clearly distinguish theory and evidence because children often distorted the evidence to match their prior beliefs.

Although these studies suggest that data itself is important in detecting causal relationships and in evaluating hypotheses, there is little evidence about the point at which children (and adults) detect covariation in a particular dataset. In fact, in her review of scientific reasoning literature Zimmerman (2000) states that "it is not clear *how large* the difference must be in order to conclude that the two events are related" (p. 115).

If students do rely on evidence to extend or modify a theory, how do they go about such a task? Students' notions of data variability may help them determine how to weigh the potential importance of different types of information. For example, data with little variability may be considered more useful in drawing a conclusion than data with greater variability *a priori*. The relevance of the data to theory may be separately evaluated.

One approach to understanding how children use evidence to extend and modify theories is to look at category induction. In a series of studies, Gutheil and Gelman (1997) presented 8- and 9-year-old children and college adults with series of category exemplars. Participants were asked whether a given property would be expected to occur in a new exemplar. The diversity and sample size of the initial sets were varied. Results suggested that children used diversity and sample size information only in combination, but were unable to use just one successfully to infer category membership. Adults, in contrast, used each property independently, as well as jointly, in inferring category membership. In these studies, however, determining that a set was homogenous or diverse relied on domain knowledge acquired outside of the experiment.

Clearly, patterns of data play a key role in scientific inference, but what characteristics of the data guide inferences about their utility? We suggest that, in drawing conclusions about comparative data, three characteristics that indicate the amount of variation in the data are key: consistency within the patterns of data (i.e., the relative sizes of the data points), the magnitude of differences (i.e., the range of each set of data) and the presence of outliers. Data that show high consistency in the direction of effects, small differences in magnitude and few outliers suggests little variability. Data that shows low consistency in the direction of effects, large differences in magnitude, and many outliers suggests more variability. This information about variability can be assessed increasingly well with a larger number of data

points, increasing the degree to which the data itself can inform an interpretation.

As a preliminary exploration of this area, we presented children and college students with sets of comparative data, and asked them to draw conclusions about differences between the sets. The data were varied systematically in number of data points presented and consistency within the pattern of data.

Method

Participants Thirty nine third graders (mean age = 9.1), seventeen sixth graders (mean age = 11.8), and fifty college undergraduates (mean age = 20.2) participated.

Procedure All participants were interviewed individually. Participants were randomly assigned to one of two conditions. The conditions differed in the cover story for the data presented. In the first condition, each participant was read the following information:

> Some engineers are testing new sports equipment. Right now, they are looking at the quality of different sports balls, like tennis balls, golf balls and baseballs. For example, when they want to find out about golf balls, they use a special robot launcher to test two balls from the same factory. They use a robot launcher because they can program the robot to launch the ball with the same amount of force each time. Sometimes they test the balls more than once. After they run the tests, they look at the results to see what they can learn.

In the second condition, we used an isomorphic background story in which two athletes were trying out for one slot on a team in different sports. The coaches asked the participants to perform certain tasks (e.g., hit a golf ball as far as possible) to assess which athlete would be better for the team. This condition was designed to see if adding information about a highly likely potential source of variability (human error) would change participants' responses in any way.

After reading the cover story, the participants were shown a series of datasets, one at a time. For each example, there was data for either two different balls of the same type, which were not given any distinguishing characteristics (e.g., "Baseball A" and "Baseball B"), or for two athletes about which there was no information other than their names (e.g., "Alan" and "Bill"). In the athlete condition, different names were used for each story, to prevent any carry-over knowledge effect. For each dataset, there were 1, 2, 4, or 6 pairs of data. Each page contained two columns of data: one listing the distance the first ball traveled and one listing the distance the second ball traveled.

The datasets varied in (a) sample size, (b) whether the datasets overlapped or not, and (c) in whether the variability in the data was high or low relative to the means. Each participant received a total of 14 comparisons, with 8 trials including no overlap (sample size 1, 2, 4, and 6), and 6 trials including one or two overlapping data points (sample size 4 with one overlapping data point, and sample size 6 with one and two overlapping data points). Half of the trials had high variability, in which the standard deviation was 15-20% of the mean, and half had low variability, in which the standard deviation was less than 2% of the mean. Each of the fourteen trials tested a different type of sports ball. (See Table 1 for specific examples of the different data characteristics.)

Table 1: Examples of datasets

Example 1: Six data pairs, no overlapping data points, low variability within columns relative to the mean, robot condition

Golf Ball A	Golf Ball B
466 feet	447 feet
449 feet	429 feet
452 feet	430 feet
465 feet	446 feet
456 feet	437 feet
448 feet	433 feet

Example 2: Four data pairs, one overlapping pair (3 out of four times Carla throws farther), high variability within columns relative to the mean, athlete condition

Carla	Diana
51 feet	38 feet
63 feet	50 feet
43 feet	56 feet
57 feet	44 feet

For each dataset, participants were first asked what the engineer or coach could find out as a result of this information and to explain any reasons for their answer. Then they were asked how sure they were about these conclusions. To answer the questions about sureness, participants were offered a four-level scale from which to select their answer, choosing among "not so sure," "kind of sure," "pretty sure," and "totally sure."

Participants were next asked if the robot or athlete launched Ball A again, exactly how far they thought the ball would go, and how sure they were that Ball A would travel X feet. They were asked the same questions about Ball B, and then they were asked how sure they were that the ball they had just named as going the farther distance would actually go farther. For example, if they said that they expected Ball A to travel 50 feet and Ball B to travel 60 feet, they were asked how sure they were that Ball B would travel farther. Again, in addition to rating their

sureness, they were asked to offer any reasons for their choices.

This series of questions was repeated for each of the 14 sets of data.

Results

Measures Participants rated how sure they were about their conclusions four times for each dataset: They gave a rating of their confidence in the initial conclusions drawn from the data, in the predictions they made of exactly how far each ball would go, and about which ball would go farther on the next trial. The ratings were assessed on a four-point scale, converted to a four-point variable, with 1 equivalent to not sure and 4 equivalent to totally sure.

In addition, participants offered reasons for their initial conclusions and final predictions of relative position. These reasons were coded for mention of any of the following factors: the precise proportion of times one ball went farther, a trend in the data, sample size, the magnitude of the difference between the two datasets, whether the datasets overlapped, a property of the ball that affected the results, a property of the robot or athlete that affected the results.

Participants also made specific numerical predictions of how far the two balls would go if launched one more time.

Levels of confidence Mixed model ANOVAs were used to examine the effects of condition, data size, overlap, and variability on ratings of sureness on the four-point scale. For each assessment, data size, overlap, and variability were within-subjects variables, and subject was treated as a random variable.

Across all age groups, there was no effect of condition (robot or athlete) on ratings of sureness. Therefore, on later analyses of these questions, we collapsed the data across conditions.

For conclusions about how sure the engineer or coach could be, based on the original data, that one athlete/ball threw/went farther, there were several notable effects. College students were highly sensitive to sample size, the sixth graders showed a small but not significant trend upwards, and the third graders showed a small but significant inverse trend. The data are summarized in Figure 1. Overall, there was a highly significant effect of grade ($F(2, 1448) = 56.38$, $p < 0.0001$), with third graders on average much more sure than sixth graders, who in turn were more sure than college students, across all sample sizes smaller than 6.

In addition to sample size, participants demonstrated a sensitivity to the presence of overlapping data points, such that they were less sure of conclusions when the data contained overlapping points. This effect was significant for all grades.

Similar patterns emerged on the assessments of participants' sureness about their predictions, both about the specific distance the balls would travel, and about which ball they expected to go farther on a repeated trial. The strongest relationships were for the college students, who appeared to always link their sureness rating to the number of data points, the proportion of overlapping data points, and occasionally to the level of variation in the data.

Although similar features in the data affected the level of sureness, there were striking differences in the sureness responses to the different questions. Participants were much less sure about specific predictions than about overall conclusions or about predictions of relative placement on a future trial. Overall, general linear models for each dataset, considering the measure of sureness as a repeated factor indicate a strong relationship between both grade and the specific question asked (i.e., the assessment of sureness for general conclusions, specific predictions and relative predictions). Across all grades, there was a relationship between the level of sureness and what question was asked, but it was weakest in the third graders and strongest in the sixth graders.

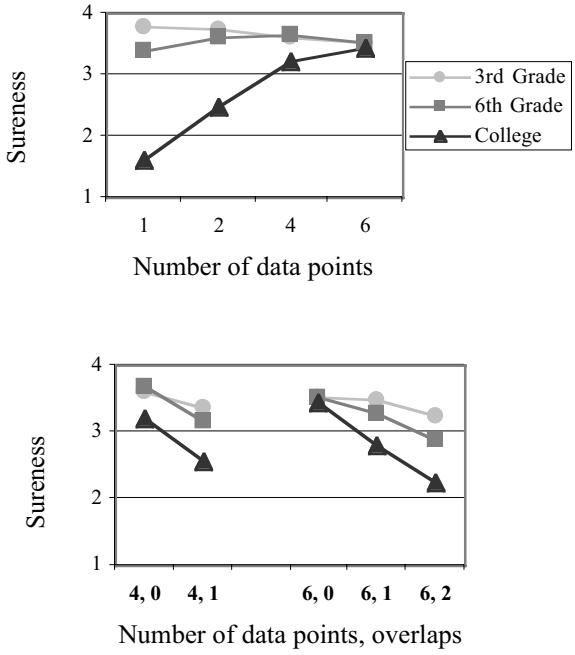

Figure 1: Average ratings of sureness by grade, sample size and number of overlapping data points

Reasons offered Participants offered justifications for confidence in their conclusions and for the predictions they made about which ball would go farther on a subsequent trial. We examined whether participants mentioned each reason at any point in response to a question about each of the fourteen datasets. We then ran Chi-square tests to look for grade differences in the frequency of participants mentioning each factor.

Table 2 presents a summary of these results. An overwhelming majority of the responses were in reference to the data and not to theoretical issues such as properties of the ball or robot/athlete. Young children were most likely to mention either the proportion of the data (e.g., "Five out of six times A went farther") or a trend in the data (e.g., "B generally went farther"). In general, college students used a much wider range of responses than younger children, with nearly all of them mentioning sample size at least once. Interestingly, despite the significant increasing trend, only a small percentage of the participants mentioned within-column variability, one of the factors we manipulated (variability in Table 2).

Table 2: Percentage of participants at each age who mentioned each justification for their sureness ratings

	3rd Gr.	6th Gr.	College
Data responses			
Precise proportion	92	86	86
Trend in data*	90	89	100
Sample size**	10	25	96
Overlap	56	64	72
Variability**	0	11	28
No Overlap**	5	7	58
Magnitude of diffs**	36	75	90
Outlier*	0	0	10
Theory responses			
Ball property	8	14	14
Robot/athlete property	18	18	22

Grade differences: *p<0.05; **p<0.01

Justifications of predictions of which ball would go farther in a new trial followed a very similar pattern as that described above, for all three age groups.

One area in which a condition difference might be expected is in use of justifications that refer to qualities of the ball, the robot, or the athlete. Mentioning the property of the robot or athlete did vary considerably by condition, with nearly all mentions in the athlete condition (i.e., participants sometimes said that a property of athlete was a reason for the outcome, but almost never attributed it to a property of the robot). This trend was even stronger when justifying predictions of future outcomes.

Predictions Participants predicted how far each ball would go if the experiment were repeated. The data from two third graders and one sixth grader were not included in this analysis because they included numbers that differed from the mean by more than twice the range of the data. These outliers skewed the data considerably, and suggested that these participants did not understand the prediction task.

Overall, however, the participants were very good at predicting how far the balls would go, and their predictions averaged close to the mean. Third graders averaged predictions that were 108.4% of the actual data means (SD =10.0); sixth graders averaged predictions that were 102.7% of the means (SD = 3.7); and college students averaged 100.1% of the means (SD = 1.4).

Discussion

Our overall conclusion is that in the absence of clear domain knowledge upon which to base theoretical explanations, children and college students paid attention to several features of data. At the same time, there were clear age differences in many responses, indicating changes over time that likely come from a combination of education, experience, and development.

In all age groups, participants were less confident about conclusions from datasets in which there were overlapping data points, indicating a sensitivity to variation in the patterns of the data. College students were significantly more confident when there were more data points, though third graders were actually slightly more confident with smaller sample sizes. It is possible that the third graders were overwhelmed by the variability and became more confused about drawing conclusions when there were more data points to consider.

Participants also showed an appreciation for some types of variability by differentiating their sureness ratings for different types of predictions. At all ages, they were more sure of conclusions about relative distances on a future trial than about specific distances, with the effect most pronounced in college students. This response pattern indicates an expectation that variation is more likely in precise measurements than in overall patterns of results. Participants seemed less attuned to within-column variability in the data, rarely citing it as a justification for either their conclusions or predictions, though older participants were still more likely to cite it.

The lack of major differences between conditions was an unexpected outcome. We had anticipated that the two different ways of framing the data would lead to different theoretical explanations of the data. However, most of the justifications offered for drawing conclusions from the data were based on the numerical evidence (e.g., a trend in the data, sample size), while very few were linked to mechanistic explanations such as a feature of the ball that might cause the outcome. There was a small but significant trend for those in the athlete condition to be more likely to justify their explanations and predictions by suggesting that the athletes may have varied in some way. However, a minority of participants at all age levels used such theoretical justifications. Some researchers have argued that children mistakenly justify conclusions that should be based on data by using their background theoretical knowledge (e.g., Kuhn, et al., 1995). In contrast, we suggest that in fact when children do not have background knowledge upon which to rely, they are likely to talk about the data in justifying conclusions.

In making distance predictions, overestimation was more common than underestimation, particularly among the youngest children. Similarly, third graders often claimed to be totally sure of conclusions they could draw after seeing only one pair of data, while college students tended to reserve their enthusiasm until seeing at least four consistent pairs of data. In general it appears that third graders often overestimate both their confidence in their ability to judge the quality of evidence, and their predictions of future performance. College students were very skilled at basing predictions of future events on the mean of observed events.

This study is a first step toward a clearer understanding of the many factors that influence the use of data in different contexts. Many other data manipulations could be examined to explore this question more thoroughly. For example, one could manipulate the size of within-column variation, the relative size of outliers, and the size of the means. In addition, one could consider multiple groups of data or a single set of data to be evaluated.

Our long-term goal is to better understand the interaction between features of data and theoretical framing, as people of all ages most often encounter data in contexts in which they have some background knowledge. This study was designed to tease apart some of the specific features of data that are used when there is not a lot of theoretical background knowledge influencing conclusions drawn. Future studies will continue to examine different characteristics of data within a range of contexts, to learn more about how they interact to affect reasoning.

Acknowledgments

This research was funded in part by grants to David Klahr from the McDonnell Foundation and from NIH. We owe many thanks to Anne Siegel and Jen Schnakenberg for assistance with data collection, data entry, and draft comments. Thanks also to David Klahr and three anonymous reviewers for comments on an earlier version of this paper.

References

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*, 1098-1120.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: The Free Press.

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

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. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237-251.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.

Klahr, D., Chen, Z., & Toth, E. E. (2001). Cognitive development and science education: Ships passing in the night or beacons of mutual illumination. In S. M. Carver & D. Klahr (Eds.) *Cognition and instruction: 25 years of progress* (pp. 75-119). Mahwah, NJ: Lawrence Erlbaum Associates.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasonin*g. Cambridge, MA: MIT Press.

Kuhn, D. (1989). Children and adults as intuitive scientist*s. Psychological Review, 96*, 674-689.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. Orlando, FL: Academic Press.

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development*, *60* (4), pp. 1-128.

Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education*, *18*, 955-968.

Masnick, A. M., & Klahr, D. (2001). Elementary school children's understanding of experimental error. In J. D. Moore & K. Stenning (Eds). *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, (pp.600-605). Mahwah, NJ: Lawrence Erlbaum Associates.

Schlottman, A. (2001). Children's probability intuitions: Understanding the expected value of complex gambles. *Child Development*, *72*, 103-122.

Shaklee, H., Holt, P., Elek, S., & Hall, L. (1988). Covariation judgment: Improving rule use among children, adolescents and adults. *Child Development*, *59*, 755-768.

Shaklee, H. & Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Development*, *52*, 317-325.

Shaklee, H., & Paszek, D. (1985). Covariation judgment: Systematic rule use in middle childhood. *Child Development, 56*, 1229-1240.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*, 99-149.