Separating signal from noise: Children's understanding of error and variability in experimental outcomes

Amy M. Masnick, David Klahr, Bradley J. Morris

Hofstra University, Carnegie Mellon University, Grand Valley State University

A young child eagerly awaits the day when she will pass the 100 cm minimum height requirement for riding on the "thriller" roller coaster at her local amusement park. She regularly measures her height on the large-scale ruler tacked to her closet door. As summer approaches, she asks her parents to measure her every week. A few weeks ago she measured 98 cm, last week 99.5 cm, but today only 99.0 cm. Disappointed and confused, when she gets to school she asks the school nurse to measure her, and is delighted to discover that her height is 100.1 cm. Success at last! But as she anticipates the upcoming annual class excursion to the amusement park, she begins to wonder: what is her real height? And more importantly, what will the measurement at the entrance to the roller coaster reveal? Why are all the measurements different, rather than the same? Because she is a really thoughtful child, she begins to speculate about whether the differences are in the thing being measured (i.e., maybe her height really doesn't increase monotonically from day to day) or the way it was measured (different people may use different techniques and measurement instruments when determining her height).

As this hypothetical scenario suggests, children often have to make decisions about data, not only in formal science classroom contexts, but also in everyday life. However, data vary. Data are imperfect both in the "real world" and in science classrooms. Learning when that variation matters and when it does not – separating the signal from the noise – is a difficult task no matter what the context. Children have two disadvantages in interpreting data. First, they

disciplines, strongly-held hypotheses require a lot of disconfirming evidence before they are revised, while those with less theoretical grounding are more easily revised so as to be consistent with the latest empirical findings.

But how does a child determine when such variation matters? As discussed above, knowledge guides interpretations of data yet data also guide the evaluation and creation of knowledge. There seem to be (at least) two plausible developmental explanations: knowledge precedes data or data precede knowledge. Although these characterizations are slightly exaggerated, it is useful to examine the implications of each. It is possible that children only begin to attend to data when they detect inconsistencies with their existing knowledge. For example, the child in our opening scenario who holds the belief that growth is a monotonic function -- and that therefore her height will always increase -- will use that "theory" to interpret any measurement indicating a "loss" of height, as inconsistent with the current theory. This anomaly may motivate a more careful and skeptical analysis of the discrepant measurement. She might look for and evaluate a series of possible explanations that account for the unexpected data. (Chinn & Brewer, 2001) Thus, through the detection of theoretical inconsistencies, children might begin to attend to data and these data in turn, provide information on the type and extent of knowledge change that is necessary.

Conversely, it is also possible that knowledge is the result of data accumulation. Perhaps

A correct test involves setting up two ramps with identical settings on every level except surface, running the test, and then measuring and interpreting the results.

We distinguish five stages in the experimentation process: design (choosing variables to test), set-up (physically preparing the experiment), execution (running the experiment), outcome measurement (assessing the outcome), and analysis (drawing conclusions). Each stage is directly associated with a different category of error.

Design error

Decisions about which factors to vary and which to control are made in the design stage. These decisions are based on both domain-general knowledge, such as how to set up an unconfounded experiment, and domain-specific knowledge, such as which variables are likely to have an effect and therefore should be controlled. Domain-specific knowledge is used to form the operational definitions of the experiment's independent and dependent variables.

Design error occurs in this stage of an experiment when some important causal variables not being tested are not controlled, resulting in a confounded experiment. Design errors occur "in the head" rather than "in the world," because they result from cognitive failures. These failures can result from either a misunderstanding of the logic of unconfounded contrasts, or inadequate domain knowledge (e.g., not considering steepness as relevant to the outcome of a ramps comparison).

Measurement error

Measurement error can occur during either the set-up stage or the outcome measurement stage. Error in the set-up stage is associated with the readings and settings involved in arranging the apparatus and calibrating instruments, and error in the outcome measurement stage is associated with operations and instruments used to assess the experimental outcomes. Measurement always

includes some error, producing values with some degree of inaccuracy. These inaccuracies can affect either the independent or the dependent variables in the experiment. Of the four types of error, measurement error most closely corresponds to the conventional view of an error term that is added to a true value of either the settings of the independent variables or the measurement of the dependent variables.

Execution error

The execution stage covers the temporal interval during which the phenomenon of interest occurs: in other words the time period when the experiment is "run." For example, in the ramps experiment, this stage lasts from when the balls are set in motion until they come to rest. Execution error occurs in this stage when something in the experimental execution influences the outcome. Execution error can be random (such that replications can average out its effects) or biased (such that the direction of influence is the same on repeated trials), and it may be obvious (such as hitting the side of the ramp) or unobserved (such as an imperfection in the ball).

Interpretation error

Although interpretation occurs during the final stage – analysis – interpretation error can be a consequence of errors occurring in earlier stages and propagated forward. That is, undetected errors in any stage of the experiment can lead to an interpretation error. For example, not noticing the ball hitting the side of the ramp as it rolls down might lead one to be more confident than warranted in drawing conclusions about the effect of the ramp design.

Even if there are no earlier errors of any importance, interpretation errors may occur in this final stage as conclusions are drawn based on the experimental outcome and prior knowledge. Interpretation errors may result from flawed reasoning strategies, including inadequate understanding of how to interpret various patterns of covariation (Amsel & Brock, 1996; Shaklee

presented with a choice between a conclusive and an inconclusive experimental test, can make the correct choice, although they cannot yet design such a conclusive test. Similarly we would expect that children might be able to recognize error-based explanations as plausible even if they are unable to generate execution or measurement error-related reasons for data variability.

Varelas (1997) examined third and fourth graders' reasoning about errors in the execution

One of our initial goals was to explore what children understand about different types of error in an experimental context. We presented elementary school children with a situation in which they had to work through each phase of an experiment: we asked them to design, execute, measure and interpret results from an experiment. At each of these stages, there was the possibility of error both in the particular phase and in the possible interpretations of the outcome (Masnick & Klahr, 2003).

We used the domain of ramps because it is a familiar domain and one that yields data with consistent main effects but some variation. We presented 29 second and 20 fourth graders (average ages 8 and 10) with the opportunity to design several experiments with ramps to determine the effects of the height and surface of the ramp on the distance a ball travels. Children were asked to make predictions, justify their designs and predictions, and run the experiment. They were then asked to draw conclusions from the results and to speculate on what the outcome would be if the experiment were to be rerun with no changes in the setup. They were asked to assess how sure they were of their conclusions on a four-point scale (totally sure, pretty sure, kind of sure, not so sure). They were also asked to generate possible reasons for variation in datasets and to reason about the effect of different factors on hypothetical outcomes.

Results

When children designed comparisons to test target variables, most trials included a number of errors in each phase of their experiments, some avoidable, others not. Children recognized some but not all of these errors and had difficulty linking their conclusions with the empirical data. In the design phase, children often made design errors, by failing to set up unconfounded experiments (16% of the second graders' designs were unconfounded; 40% of the fourth graders' were). However, their justifications for their designs and their outcome predictions were

associated with the accuracy of their design. In other words, participants who designed confounded experiments were likely to expect all the variables they did contrast to affect the outcome, even when that was not the stated goal of the comparison. Similarly, those children who did not vary the target variable were much less likely to cite differences in the target variable as a justification for the expected outcome. This finding suggests an understanding of the causal link between the design and outcome, even when this link was not clearly articulated.

In measuring the distance a ball traveled, the likelihood of measurement error was small due to the constrained nature of measurement in this task: the distance balls rolled was measured discretely by noting the numbered step on which the ball landed. However, nearly all of the participants were able to name sources of measurement error when asked to explain variation in

All of the data sets varied slightly: there were no two measurements given to participants that were identical within the set of trials for each given variable. However, because string length was the only variable that caused a true effect, the difference in the readings from the trials with a short string and those with a long string were much more pronounced. The data from these two sets of trials did not overlap. Figure 1 shows the box-plots for the data that were presented to each participant. (Note that the participants received the data one data point at a time, after timing each round of swings. The data were recorded in a column format, on a preprinted handout provided by the experimenter.)

As noted above, each participant experimented with the effects of length, weight, and height. Because we wanted examine participants' ability to "calibrate" high vs. low variability in this context, we presented the length variation as the first factor to be explored for one half of the participants, and as the last factor to be explored for the other half.

Results

Both adults and children learned from running the experiments, and there was no effect of whether the length variable was presented first or last. Figure 2 shows several very clear patterns: (a) With respect to initial knowledge, both adults and children tended to believe (correctly) that length matters, and (incorrectly) that weight and height also matter. (b) Both adults and children learned from pre-test to test phase about all three variables. (c) Adults not only knew more about all three factors than did children at pre-test but also improved their knowledge more than children after seeing the data. In other words, by the end of the test phase, and through the post-test phase, most adults had revised their faulty beliefs about the effect of height and weight on the period of a pendulum. In contrast, children's gains, although statistically significant, remained at very low levels. These differences cannot be attributed to

how children evaluate the data itself. That is, how do children decide when evidence is compelling or uncompelling? Once these characteristics of data are identified, we can examine the extent to which specific characteristics of data are related to theoretical change. To examine this issue, a different approach was used in which data were presented with little theoretical guidance so that the conclusions drawn would be predominantly data-driven rather than theory-driven. In this study, we examined children's reasoning in the interpretation phase of an experiment: data were presented as results of a completed experiment, and participants were asked to draw conclusions based on the information they had available. We set up situations with minimal theoretical background information, to make the variation in data characteristics particularly salient.

Two of the most important ideas about data involve expectations about sample size and expectations about variation in data distribution , and we used these variables as the focal variables in our study. We asked participants to draw conclusions about whether there was a difference between two sets of data and to explain their reasoning (Masnick & Morris, 2002).

Thirty nine third graders, forty-four sixth graders, and fifty college undergraduates were presented with a cover story, and then were asked to reason about potential differences between two sets of data. Half of the participants read the following cover story about engineers who are testing new sports equipment, using robot launchers to repeatedly test different sports balls, such as tennis balls and golf balls. The other participants read an isomorphic cover story about a coach trying out two athletes vying for one slot on her team.

> 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.

After reading the cover story, participants were shown a series of datasets, one at a time. For each example, there were 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

Participants were also asked for justifications for their reasoning about why they felt they could be sure of conclusions. Coding of data-based reasons involved noting whether participants mentioned a trend in the data ("5 out of 6 times A went farther"); sample size ("It's only two times so it's hard to tell"); no overlap ("A always went farther than B"), variability within the column ("the numbers were really far apart in A," and magnitude of differences ("A went a lot farther than B")). There were large grade differences in the frequency of using each of the descriptions. See Table 3. However, all but one participant (a third grader) made at least one explicit reference to data characteristics such as a pattern in the data or the magnitude of differences. This finding indicates that even as early as third grade, children are paying attention to some characteristics of data, and using this information in guiding their conclusions.

Insert Table 3 about here

Additionally, reasons were also coded to take note of whether they included a mechanistic explanation for the outcome. These responses were classified as a reason based on a property of the ball ("Ball A was more aerodynamic"), of the robot or athlete ("Maybe the robot was breaking down when it threw Ball B"; "Bill was getting tired"), or of the environment ("Maybe there was wind when Ball A was thrown"). Mentioning the property of the robot or athlete was the only factor we found that 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 very rarely attributed it a property of the robot.) However, there were no grade differences in frequency of providing a mechanistic explanation, with an average of 50% of participants providing at least one mechanistic explanation. In their interpretations, a sizeable number of participants were using prior background knowledge to explain the data.

Discussion

References

Amsel, E., & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development, 11,* 523-550.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data, *Cognition and Instruction, 19*, 323-393.

Chinn, C. A., & Malhotra, B. A. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology*, *94*, 327-343

delMas, R. & Liu, Y. (this volume) Students' conceptual understanding of the standard deviation. In M. Lovett & P. Shah (Eds.) *Thinking with Data: The 33rd Carnegie Symposium on Cognition.*

Garfield, J., delMas, R. & Chance, B. (this volume). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M. Lovett & P. Shah (Eds.) *Thinking with Data: The 33rd Carnegie Symposium on Cognition.*

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.

Hon, G. (1989). Towards a typology of experimental errors: An epistemological view. *Studies in History and Philosophy of Science, 20*, 469-504.

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

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language, 32,* 402-420.

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*, 748-769.

Konold, C. (1991). Information conceptions of probability. *Cognition and Instruction*, *6*, 59-98.

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

Krajcik, J. & McNeill, K. (this volume). Middle school students' use of evidence and reasoning ..in writing scientific explanations. In

Metz, K. E. (1998). Emergent understanding and attribution of randomness: Comparative analysis of the reasoning of primary grade children and undergraduates. *Cognition and Instruction, 16,* 285-365.

NRC (2000). Inquiry and the National Science Education Standards: A guide for teaching and learning. Washington, DC: National Academy Press.

From “Error matters: An initial exploration of elementary school children’s understanding of experimental error” by Masnick & Klah 94Daf

Table 2: Examples of datasets shown to participants

Example 1: Six data pairs, no overlapping data points, 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), athlete condition

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

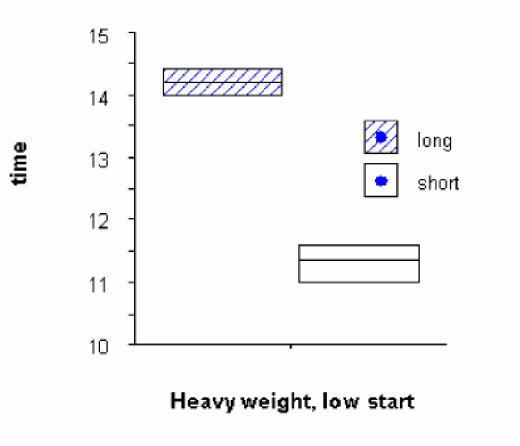
Table 3 Percentage of participants at each grade level who gave each data-based explanation at least one time.

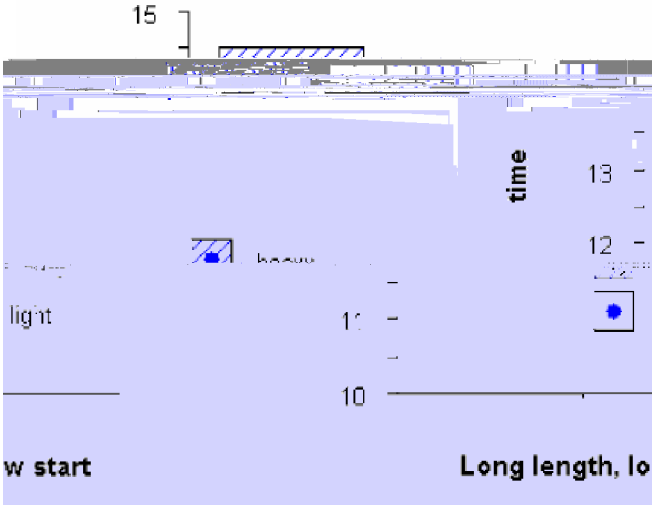
	3rd grade	6th grade	Undergraduate
Trend	90	84	100
Sample size	10	27	96
Overlap	56	61	72
Variability	0	7	28
Magnitude of difference	36	80	90

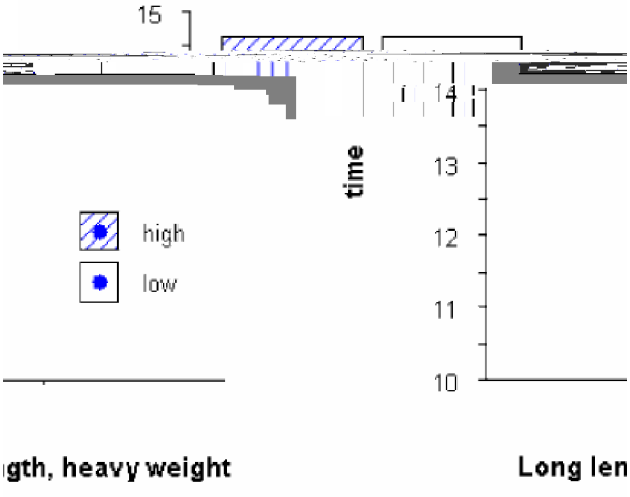
Figure captions

Figure 1. Summary of data presented to participants, showing distinct separation between times associated with long vs. short strings, and complete overlap for times associated with heavy/light weights and high/low starting positions. (Y-axis shows seconds to complete 10 swings; box plots are based on four data points for each sub-plot.)

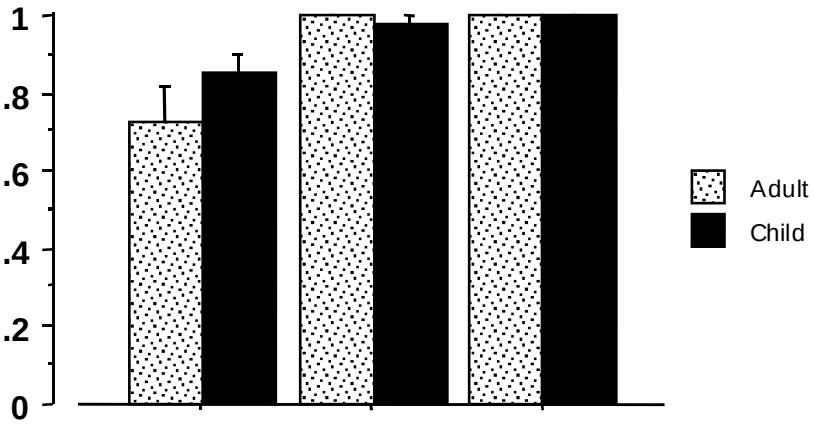
Figure 2 The percent of participants in each age group who believed each variable made a


15
14
13
12
11
10
time
long
short
Heavy weight, low start

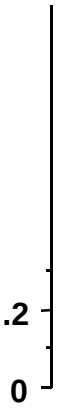

15
time
13
12
10
Long length, lo

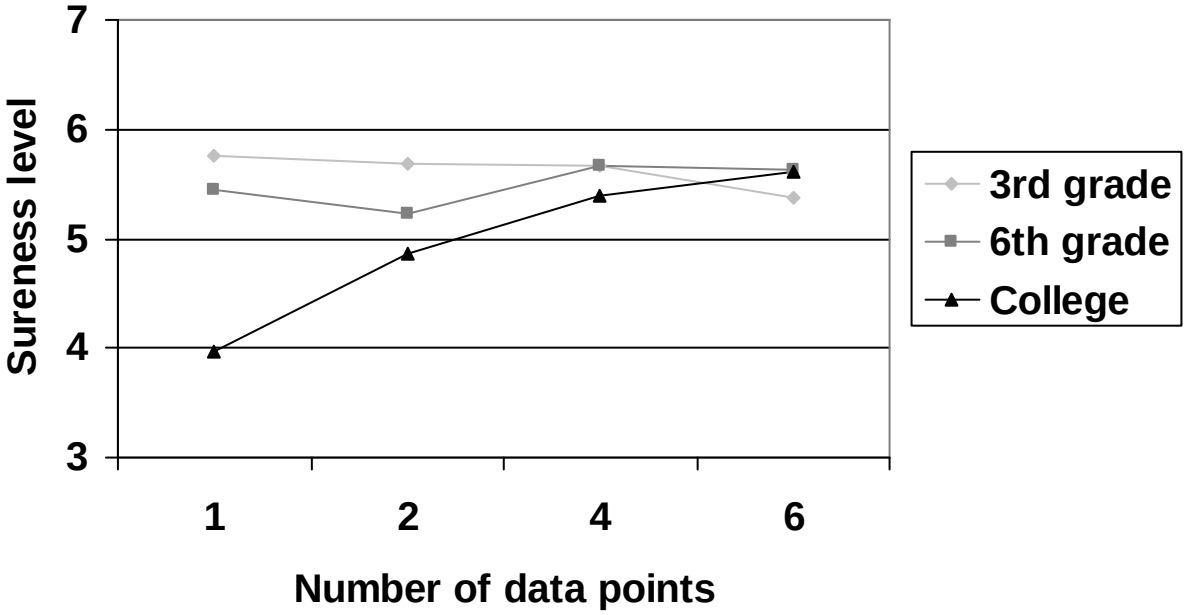

15
time
13
12
11
10
high
low
Long len

2A



2B




7
6
5
4
3
Sureness level
1
2
4
6
Number of data points
3rd grade
6th grade
College