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 have no formal statistical knowledge, which makes it impossible for them to fully assess the properties of the data in question. Second, children’s limited experience makes it difficult for them to detect data patterns and to formulate coherent expectations – based on nascent theories – about natural phenomena.

In contrast, adults with formal statistical training can use those tools in the science laboratory to distinguish real effects from error, or effects caused by factors other than the ones being explored. When statistical tools reveal that observed differences are highly unlikely to have occured by chance, those with statistical training can feel more confident about drawing conclusions from data. Another critical component to such reasoning is theory, which we define as the background knowledge and experience brought to the task, that influences decisions about the importance of variability and the reasonableness of the conclusions. This theoretical context may include hypotheses about potential mechanisms that lead to observed outcomes, but may also be a simple statement that events are related or that they do not contradict explanations of other phenomena. The theoretical component involves any claims about the data, based on information other than the data themselves.

In everyday reasoning, for those with or without statistics training, deeply held beliefs require large and consistent discrepancies between expected outcomes and empirical discrepancies before theory revision can take place. From a Bayesian perspective, the impact of new evidence is modulated by the size of the prior probabilities. For example, if a person has seen 1000 instances of events x and y co-occurring, one instance of x occurring by itself is unlikely to change the expectation that the next instance of x is likely to be paired with y. And even in the classic Fisherian statistical world of t-tests and ANOVAs, the significance of statistical results is always tempered by a theory-embedded interpretation.   In all scientific 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 children detect patterns in their environment and use the data as the basis for conceptual groupings. For example, it has been suggested that several facets of language acquisition (e.g., phoneme tuning) are derived from the statistical structure of the child’s language environment. In phoneme learning, once the acoustical properties of a set of phonemes have been derived, children prefer these sounds to novel sounds (Jusczyk, Friederici, Wessels, Svenkerud. &

Jusczyk, 1993) Once established, these conceptual units might anchor expectations about the probability of occurrences in the environment. A possible consequence of such an expectation is that deviations from the established data patterns (e.g., statistical structure) provide evidence of variation and may require changes in current knowledge.

We argue that theoretical, background knowledge and data dynamically interact in reasoning. That is, the tendency to attend to theoretical claims and explanations, or to specific data will be driven by the degree to which each element matches (or does not match) current knowledge. Many researchers have noted that theory influences the interpretation of data (e.g., Chinn & Malhotra, 2002; Koslowski, 1996; Kuhn & Dean, 2004; Schauble, 1996). For example, people often discount data that contradict their current knowledge. Yet few researchers have examined the role of data in modifying theory. Further, there has been little research in which the interaction between reasoning about information based on prior knowledge and reasoning explicitly about data characteristics have been examined.

In this chapter, we examine the relative roles of reasoning context and data characteristics when children and adults reason about error. First, we describe the framework that has guided our research in this area and then we discuss three empirical studies of these issues.

Beliefs based on background knowledge or context are one key component in drawing conclusions. However, another important, often-overlooked component in decision-making involves beliefs about characteristics of data. In the science laboratory, these beliefs are usually formalized in the use of statistics, but for children, these beliefs are based on a basic knowledge of probability and informal notions of statistics.

National science education standards urge teachers to develop children’s critical thinking skills that include "deciding what evidence should be used and accounting for anomalous data" (NRC, 2000, p. 159). In conducting authentic scientific investigations, students encounter variability and error in data collection and need to discriminate meaningful effects from experimental errors. In addition, to "account for anomalous data," students need to understand the various sources of error (e.g., procedural errors, uncontrolled variables, experimenter bias). This knowledge provides a foundation for evaluating evidence and drawing conclusions based on scientific data. Thus, it is not sufficient (though necessary) for students to be able to analyze data using mathematical and statistical procedures. Rather, it is essential for students to be able to reason meaningfully and coherently about data variability itself.

Error taxonomy

One way to consider data use in science education is to consider data in the context of an experiment, noting that data variation can occur in any of a series of places. Variation and the interpretation of the variation can have different consequences at different stages of an experiment. To structure the approach to looking at these issues, we developed a taxonomy of types of errors, considering errors by the phase of experimentation during which they occur (Masnick & Klahr, 2003), building on Hon’s earlier epistemological taxonomy of error (1989). The taxonomy is summarized in Table 1.

Insert Table 1 about here.

The taxonomy identifies five stages of the experimentation process and four types of error that can occur during these stages. Our description is couched in terms of a simple ramps experiment in which participants are asked to set up two ramps such that they can be used to test the effect of a particular variable, such as the surface of the ramp, on the distance a ball travels.

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

& Paszek, 1985) or from faulty domain knowledge that includes incorrect causal mechanisms (Koslowski, 1996). Both statistical and cognitive inadequacies in this stage can result in what are conventionally labeled as Type I or Type II errors, that is, ascribing an effect when in fact there is none, or claiming a null effect when one actually exists.

Operationally, the assessment of interpretation errors must involve assessing both the conclusions drawn and one's confidence in the conclusions. Sometimes this assessment is defined formally by considering whether a statistical test yields a value indicating how likely it is that the data distribution could have occurred by chance. Regardless of whether statistics are used, a final decision must be reached about (a) what conclusions can be drawn, and (b) the level of confidence appropriate to these conclusions.

Background literature

Past research about children's understanding of experimental error and data variability has come from a range of contexts and methodologies. Some researchers have examined understanding of probability while others have looked at understanding in classroom contexts. Here we briefly summarize the related research.

Early research into children's understanding of data examined their conceptions of probability (Piaget & Inhelder, 1951/1975). Piaget and Inhelder suggested that children under the age of ten have difficulty with large sample sizes, often becoming more confused with increased amounts of data.

Given the complexity of a concept of experimental error, it is likely that children master different components of it along different developmental (and educational) trajectories. Indeed, the literature provides some evidence for such piecemeal growth of design error understanding. For example, Sodian, Zaitchik, and Carey (1991) demonstrated that even first graders, when 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 and outcome measurement stages, by looking at how they reasoned about repeated measurements. She found that most children expected some variability in measurements, although why they expected this variability was not always clear. Children also exhibited a range of opinions regarding the value of repeated measurements, with some believing the practice informative, and others finding it confusing and a bad idea. Many children appeared to believe that uncontrolled measurement and execution errors could affect outcomes, but they were often unable to explain the link between these error sources and the ensuing variation in data.

Schauble (1996) examined the performance of fifth graders, sixth graders, and non-college adults on two different tasks in which the participants' goal was to determine the influence of various factors. One difficulty many children (and some adults) had was in distinguishing variation due to errors in measuring the results and variation due to true differences between the conditions (that is, between intended contrasts and measurement stage errors). When in doubt, participants tended to fall back on their prior theories. If they expected a variable to have an effect, they interpreted variability as a true effect. If they did not expect a variable to have an effect, they were more likely to interpret the variability as due to error. Thus, their prior beliefs sometimes led them to make interpretation errors in drawing conclusions.

In more recent work, Petrosino, Lehrer, and Schauble (2003) explored fourth graders' understanding of data variability when they take repeated measurements in different contexts.

They focused primarily on what we refer to as measurement errors, and were able to teach students to think about measurements as representative of a sample of measures. They had participants use instruments with varying levels of precision, and focused discussion on the best ways to summarize the data they collected. Students trained in this way performed significantly above the national average on assessments of how to collect, organize, read, represent and interpret data.

Some researchers studying the phenomena of categorical induction have explored children's use of data characteristics from a different perspective, looking at children's ability to make inductions based on prior information. Gutheil and Gelman (1997) explored how children use sample size information and information about variation within categories in induction. They found that 8- and 9-year-old children were able to use diversity and sample size together in inferring whether a given property would be expected to occur in a new exemplar. Similarly, 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.

At the same time, there is some evidence that 11-year-old children still struggle to understand the reasons for taking 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.

Taken as a whole, this small collection of studies leaves many important questions unanswered and does not provide a coherent picture of the way that children develop an understanding of error and data variability. There is evidence of skill at some types of error-based reasoning as early as first grade, yet also evidence of difficulty in reasoning about error into adulthood. Metz (1998) has noted that the limited literature on children's ability to interpret data variability reveals an apparent inconsistency between studies suggesting that children understand probability and statistics at a young age, and those suggesting that adults have significant difficulties with these concepts (e.g., Konold, 1991; Tversky & Kahneman, 1974). Metz argues that the difference may be due to the widely discrepant criteria for precisely what behaviors do and do not indicate a solid grasp of these concepts.

Some work done by contributors to this volume also bears relevance on these issues. For instance, Krajcik and McNeill (this volume) looked at how middle school children’s reasoning changes with different kinds of content knowledge. In addition, Garfield, delMas, and Chance (this volume), and delMas and Liu (this volume) describe some of the difficulties college students have in understanding variation and the concept of a standard deviation.

In this chapter, we describe a series of studies in which we have begun a systematic, theoretically-guided examination of children’s understanding of error and data variability. Our aim is to learn more of what children know and understand about data and the ways in which they can use characteristics of data in drawing conclusions. Thus far, we have explored these issues in three contexts that vary in the extent to which children's a-priori causal theories are correct. In one context (balls rolling down ramps) most 2nd to 4th graders’ initial theories are correct, at least about most of the causal factors. In the second context (pendulums) their causal theories are mainly incorrect. In the third context (presenting outcome data for comparison), there is little theoretical context to rely on, and conclusions must be drawn solely from data.

Children’s use of data in a well-understood context: ramps

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 data, indicating a recognition that this type of error is important to consider.

In addition, children were skilled at naming and recognizing many sources of execution-stage error across different tasks in the study, citing potential issues such as positioning of the ball at the start of the ramp, method of releasing the gate, or wind in the room. This finding indicates that they understand the idea that many variables can play a role in a final outcome. When asked explicitly about error, children found it easy to propose different possibilities. In fact, 90% of participants were able to generate at least one potential explanation for data variation throughout the study (the five children who could not name any sources were second graders). However, it is not clear if children link this understanding with their other knowledge about error. For example, children were unlikely to mention execution error (or to note its absence explicitly) in justifying their level of confidence that the target variable had an effect. Thus, although there is evidence for some understanding of the role of execution error, it may not yet be integrated fully into the child’s knowledge base.

Children’s interpretation errors were assessed by their confidence in and justifications for their conclusions about experimental outcomes. Interpretation errors are the most complex type of error because correct interpretation requires integration of all available sources of information, including information from prior theories, from empirical evidence, and from knowledge of all other potential errors that could influence the outcome. Second and fourth graders’ understanding of the role of different factors in interpretation seems to be weak at best. Children were more confident about their conclusions when the data matched their prior beliefs (their predictions) than when it did not. However, they still said they were sure of the conclusions drawn from later outcomes for 76% of their incorrect predictions, compared to 95% of their correct predictions. This difference suggests that at least some children are sensitive to conflicts between theory and data. In addition, prior domain knowledge appeared to guide their reasoning; children justified most of their predictions by referring to the expected effects of the target and non-target variables.

Discussion

Overall, children were highly skilled at explaining their designs and in generating explanations of sources of variability, but these skills did not translate into an integrated approach that considered error at each stage of an experiment. They were not surprised by variation in the data and did not seem to be thrown off by it, but they had difficulty in using this information explicitly in drawing final conclusions about experimental outcomes.

In the ramps domain, children’s prior beliefs about the effect of height and surface were almost always confirmed by the data (when the experiments were designed properly). Working in this domain allowed for an important beginning study of children’s understanding of the phases of error and data variation, through examining children’s reasoning through all phases of experimentation. Presumably, they all had some knowledge of the mechanics of ramps *before* they began these experiments. This knowledge may have affected their use of data – despite the variability in specific outcomes, the overall findings were almost always in the direction children expected. The data varied but still led to the same conclusions. Thus, the causes of that variability seem likely to be due to something other than the experimental manipulation.

In contrast, when children's prior beliefs are inaccurate, they have in some ways a more difficult decision to make: when is the variation due to a true effect, and when is it due to assorted errors such as measurement or execution errors? In such a situation, the data provide new information, and do not merely support existing beliefs.

Children's use of data in a poorly-understood context: pendulums

Pendulums provide a potentially revealing context in which to explore children's understanding of data variability. Most children and adults believe (correctly) that the length of a pendulum affects its period of oscillation. However, most children and adults (!) also believe (incorrectly) that that the mass of the bob, the height from which it is released and the force with which it is released all influence a pendulum's period.

Of course, length is the sole causal factor only in an idealized situation in which no extraneous forces act on the pendulum (such as flexibility in the string or the pendulum frame, air resistance, friction in the pivot point of the pendulum, and so on). Thus, in timing the period of a pendulum in a research laboratory, any of these factors can lead to small errors and variation in the data. We were interested in how the largely incorrect beliefs of children (4th and 5th graders) and adults (college undergraduates) would influence their interpretation of data sets containing some error variation, and conversely, how variability in the data from repeated experiments would influence their beliefs.

All participants (28 undergraduates and 49 children) were interviewed individually in the lab, and their beliefs about what factors influence the period of a pendulum swing were assessed at three points: (a) in the pretest phase, before they ran any experiments, (b) during the test phase, immediately after they had "generated" and recorded the set of data points for each of the possible factors, and (c) in the posttest phase following all the experimental runs, when participants were asked to review the results of their experiments and state their final beliefs.

During phase b, participants executed a series of experimenter-designed experiments in which the students timed the period of pendulums having different string lengths, bob masses, and starting heights. For any particular configuration of these three potentially causal variables, we chose one to vary while the other two were held constant. Participants were asked to time ten swings of a pendulum in each of two configurations, such as 10 swings with a heavy bob and ten swings with a light bob, while string length and release height were held constant. The participant released the pendulum, counted ten swings, and the experimenter read out the resulting time. For each variable under investigation, this procedure was repeated 8 times: 4 for one level of the variable (e.g., a long string) and 8 for the other level (e.g., a short string). Unbeknownst to the participant, all of the times read by the experimenter were predetermined and not actual measures of the current trials. The times that we falsely presented as veridical were very close to the actual times for each pendulum trial, so that they were entirely plausible. However, this manipulation ensured that each participant was exposed to exactly the same amount of data variation. (Debriefings revealed that none of the participants suspected our deceptive manipulation.)

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 adults' being less sure of their knowledge at the outset, and thus being more responsive to data. In fact, the sureness ratings for children and adults started out at almost identical levels. By the post-test, adults were significantly more confident than they were at pre-test. They were also more confident than the children were at post-test (though their confidence increased as well).

Insert Figure 2 about here

Discussion

The results of this experiment indicate that adults are much better than children at using data to revise their beliefs when the data do not agree with prior expectations. Adults clearly differentiated the small variation in measurement of factors that do not play a role in the outcome (different weights and heights) from the large variation evident in the measurement of the one variable that did make a difference (length of the string). Children appear to have more difficulty with this process. One possible explanation is that the data variation in the non-causal variables, even though much smaller than the difference in the two levels of the causal variable, enabled them to retain their faulty beliefs. Recent work has also shown that children have much more difficulty changing inaccurate beliefs that a non-causal variable is causal than the reverse (Kanari & Millar, 2004). In the current study, the designs were set for participants, but children had difficulty in interpreting potential measurement and execution errors as causes for variability, and thus made many interpretation errors when data contradicted their beliefs. Using data to alter theory was a challenge.

Children's use of data with no theoretical context

One method of exploring the influence of theory on conclusions drawn about data is to examine reasoning in different theoretical contexts. While much research has examined children's understanding of data within theoretical contexts, there has not been much study of 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 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 dataset, to prevent any carry-over knowledge effect. Each page contained two lists of data: one listed the distance the first ball traveled and the second listed the distance the second ball traveled.

The datasets varied in sample size and in variability. Sample size was operationalized by the number of pairs of data presented with each story, with either 1, 2, 4 or 6 pairs of data presented. Variability was operationalized by varying whether the data in the two columns overlapped or not. In some cases (“no overlap”), all the values in one column were larger than all values in the other column. In other cases, the majority of values in one column were larger than values in the other column, but there were one or two pairs in which the pattern was reversed. Each participant received a total of 14 comparisons, with 8 trials including no overlap (two each of 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 or two overlapping data points). Each of the fourteen trials tested a different type of sports ball. Examples of the datasets presented are shown in Table 2.

Insert Table 2 about here

For each dataset, participants were asked what the engineer or coach could find out as a result of this information and to explain reasons for their answer. They were asked how sure they were about these conclusions, using the same four-level scale as in studies described earlier (totally sure, pretty sure, kind of sure, not so sure).

Results

We used participants' sureness as the dependent measure, and looked at the effect of different sample sizes and levels of variability. There were no differences between the robot and athlete conditions on these two factors, and so data were collapsed for these analyses. When comparing across datasets in which there were no overlapping data points, we found that college students were much more sure with more data, sixth graders had no significant differences across sample size, and third graders were actually more sure with less data. See Figure 3. Although this effect did not reach significance, as most third graders were highly confident regardless of the amount of data, it is an intriguing finding suggesting children at this age are evaluating data by very different standards than adults .

Insert Figure 3 about here

When comparing datasets with the same sample size but different levels of variability, we found that this variation in the data affected all participants' ratings of sureness. At all grade levels, participants were more sure of conclusions when the data in the two columns did not overlap than when it did. There was a linear relationship: participants were less sure of the outcome when there was one overlap than when there was none, and even less sure with two overlaps than with one.

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

These data demonstrate 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. Third graders and college students responded to sample size as a key determinant of how sure they could be in their conclusions, although surprisingly, third graders were on average less sure of conclusions with more data points. At all ages, participants responded to variation in the data as an important factor to consider in drawing conclusions.

In addition, children regularly referred to characteristics of the data in explaining their reasoning. All children noticed these characteristics, and talked about them to some degree. This finding indicates that children are aware that data vary and that this variation is something to consider in reasoning. Although they did not always use the information as they described it, the fact that they would offer data-based more often than mechanism-based explanations for the results indicates that they may be using characteristics of data to develop theoretical explanations. Such interpretations indicate an understanding that many factors at different stages of the testing phase could have an impact on the outcome and are important to consider in drawing conclusions.

At the same time, there were clear age differences in many responses, indicating changes over time. The use of sample size information changed most dramatically with time, with the slight tendency for third graders to be more sure with less data changing to a fairly consistent sureness across sample size for sixth graders and finally, a clear pattern of increased confidence by college. This finding replicates similar findings indicating that younger children are often confused with larger quantities of data (e.g., Piaget & Inhelder, 1951/1975; Varelas, 1997).

The findings from this study suggest that students pay attention to data characteristics as early as third grade, and are able to use this information in drawing conclusions. Their knowledge of experimental processes appears to be guiding their reasoning about how to interpret experimental results. Finally, in justifying their reasoning, children and adults are relying primarily on explanations based in the data, even if these are not always fully articulated or accurately applied.

Conclusions

The results from these three studies indicate that children recognize data variation in many experimental contexts, and they use characteristics of data in their reasoning. However, the context is important. When reasoning about data in a well-understood context -- when experimental outcomes match expectations -- children are unlikely to change any theoretical understanding. In such contexts, they can reason to some degree about causal mechanisms that lead to variation in data. The ramps study indicates that children expect error and variation in data, and can generate many potential causes for the sources of that error. They often make design errors, but they also recognize execution and measurement errors during all phases of experimentation.

In a poorly understood context – when experimental outcomes do not match expectations – children demonstrate difficulty integrating new data that contradict their beliefs. Children seemed to be aware of the possible sources of error but had difficulty in interpreting the differences in variability as errors and not meaningful indications of a true difference between conditions.

Finally, when presented with data without a strong theoretical context, the data study indicates that children pay attention to characteristics of data and use sample size and variability as information to consider in reasoning with data. They use the key features of data in drawing conclusions, even when they cannot always articulate their justifications. However, nearly all participants talked at some point about characteristics of the data guiding their reasoning. About half the participants also develop theoretical explanations to explain the data, speculating on potential design and execution errors that may have caused the variation.

Although we have found evidence that children do use characteristics of data in their reasoning, adults are much more skilled at the integration of theory and data. The mechanisms of this developmental shift are something it seems premature to speculate about. It is clear that experience and formal education increase children's exposure to varied data. However, we are just beginning to understand what children do know about data and how they use it. We do not have data on the frequency of children's experiences with data in real life or the classroom, and how often they are asked to draw conclusions in those situations.

To return to our original question, how do children and adults separate signal from noise? In considering the somewhat exaggerated question of which comes first, the data or the theory, we are left with evidence supporting a true bootstrapping pattern, in which data and theory knowledge work together. In the ramps study, children's (mostly accurate) theories guided their reasoning – the mechanistic explanations were the most common justifications for their conclusions, regardless of the effectiveness of the design used. In the pendulum study, adults were able to use data to update their inaccurate theories without much difficulty, but only a small number of children were able to do the same. Strong theories are difficult to overcome with a brief exposure to data. In the data study, a small but sizeable number of participants of all ages used data characteristics to generate theoretical explanations for data patterns, without specific prompting to do so. This study demonstrated that variation is a key features that children and adults attend to, while the use of sample size information is more complicated. Although adults are better able to modify theories in response to contradictory data, children show some evidence of using varied data characteristics in their reasoning and using the information both to draw conclusions and to develop causal explanations. Thus, it seems that both theory and data are important in designating what constitutes “signal” as well as “noise.”

In light of these findings, creating effective classroom materials to aid children in learning about data and experimentation needs to build on children’s nascent understanding of data variation. The fact that young children do talk about data characteristics and use some data features in drawing conclusions indicates some preliminary understanding of the nature of science experimentation. The depth of this knowledge is incomplete, and future research into its limits will aid in this goal.

So what will the young roller coaster enthusiast from our opening example conclude about her true height? That will depend in part on her age and experience. Most likely if she is in second or third grade, she will find multiple measurements confusing, and will have difficulty assessing her true height. However, because height measurement is probably a familiar context, she will have at the ready some potential theoretical explanations for the variability, considering possible measurement errors or other factors that may have gone awry. She will have difficulty integrating these theoretical beliefs with the data she sees, but she is aware that there can be variability in measurement. With increased age and experience, she will come to be more confident with increased data and to develop a more sophisticated repertoire of potential explanations for variability.

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 M. Lovett & P. Shah (Eds.) *Thinking with Data: The 33rd Carnegie Symposium on Cognition.*

Kuzmak, S. & Gelman, R. (1986). Children's understanding of random phenomena. *Child Development*, *57*, 559-566.

Kuhn, D. & Dean, D., Jr. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development*, *5*, 261-288.

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. (2003). Error matters: An initial exploration of elementary school children's understanding of experimental error. *Journal of Cognition and Development*, *4*, 67-98.

Masnick, A. M., & Morris, B. J. (2002). Reasoning from data: The effect of sample size and variability on children's and adults' conclusions. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 643-648). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Petrosino, A., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning, 5*, 131-156.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children* (L. Lowell,, Jr., P. Burrell, & H. D. Fishbein, Trans.). New York: W. W. Norton. (Original work published 1951).

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*, 102-119.

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

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*, 753-766.

Varelas, M. (1997). Third and fourth graders' conceptions of repeated trials and best representatives in science experiments. *Journal of Research in Science Teaching*, *34*, 853-872.

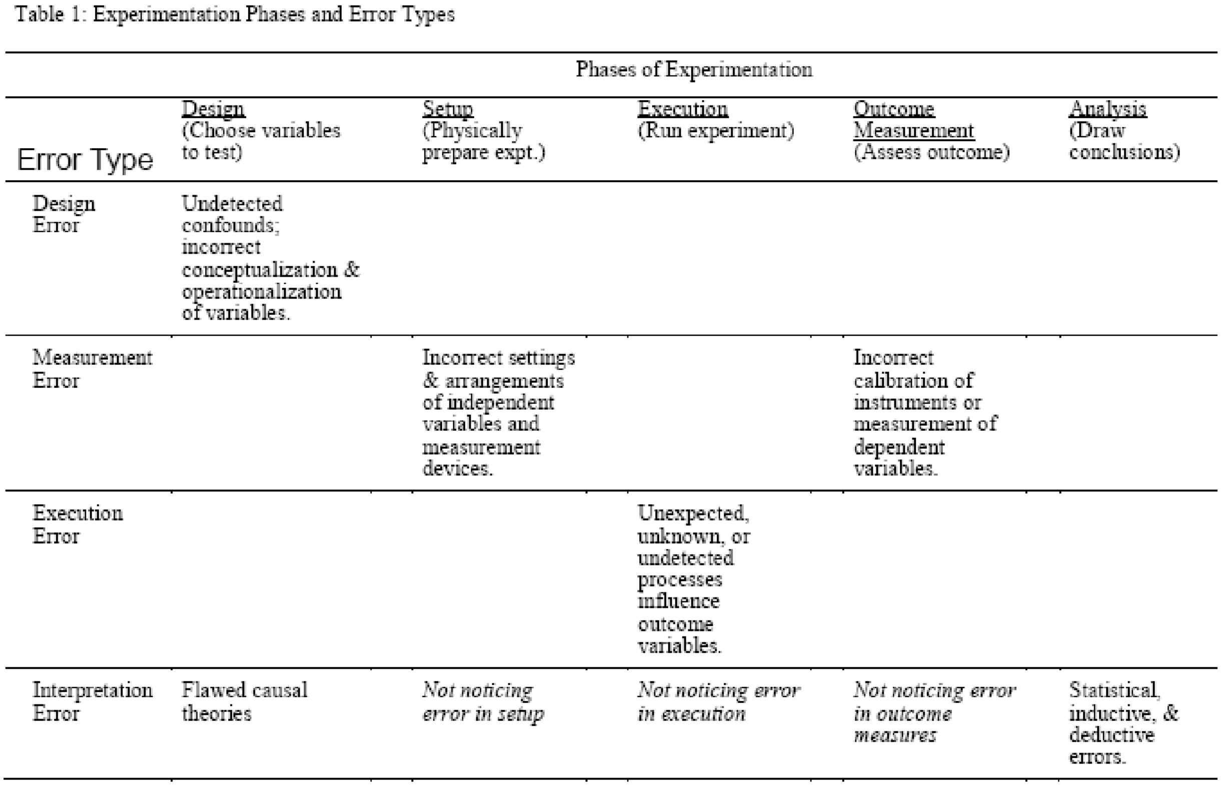


Table 1: Experimentation Phases and Error Types

	Phases of Experimentation				
Error Type	Design (Choose variables to test)	Setup (Physically prepare expt.)	Execution (Run experiment)	Outcome Measurement (Assess outcome)	Analysis (Draw conclusions)
Design Error	Undetected confounds; incorrect conceptualization & operationalization of variables.				
Measurement Error		Incorrect settings & arrangements of independent variables and measurement devices.		Incorrect calibration of instruments or measurement of dependent variables.	
Execution Error			Unexpected, unknown, or undetected processes influence outcome variables.		
Interpretation Error	Flawed causal theories	*Not noticing error in setup*	*Not noticing error in execution*	*Not noticing error in outcome measures*	Statistical, inductive, & deductive errors.

From "Error matters: An initial exploration of elementary school children's understanding of experimental error" by Masnick & Klahr, *Journal of Cognition and Development, 4*, p.70. Copyright 2003. Lawrence Erlbaum Associates. Reprinted with permission.

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.

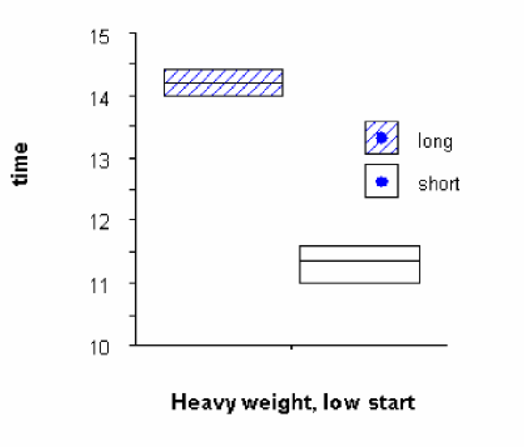
	3rd grade	6th grade	Undergraduate
Trend	90	84	100
Sample size	10	27	96
Overlap	56	61	72
Variability	0	7	28
Magnitude of difference	36	80	90

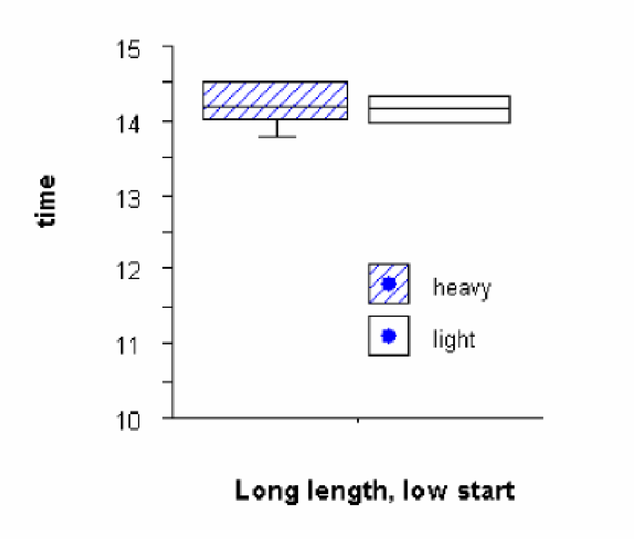
Figure captions

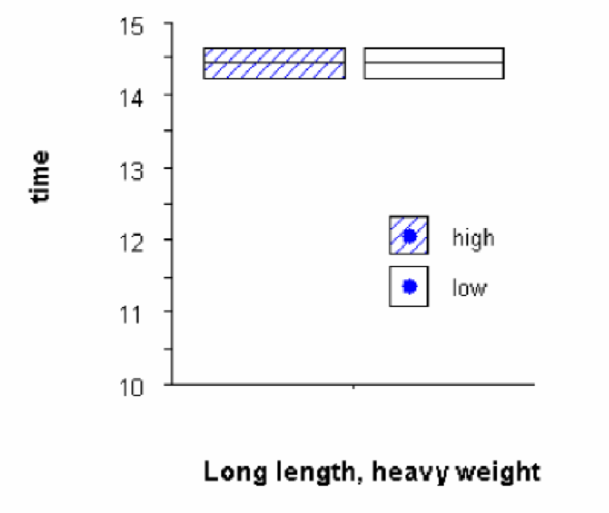
Figure 1. Summary of data presented to participants, showing distinct separation between times associated with long vs. short strings, and complete overlap for times associated with heavy/light weights and high/low starting positions. (Y-axis shows seconds to complete 10 swings; box plots are based on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

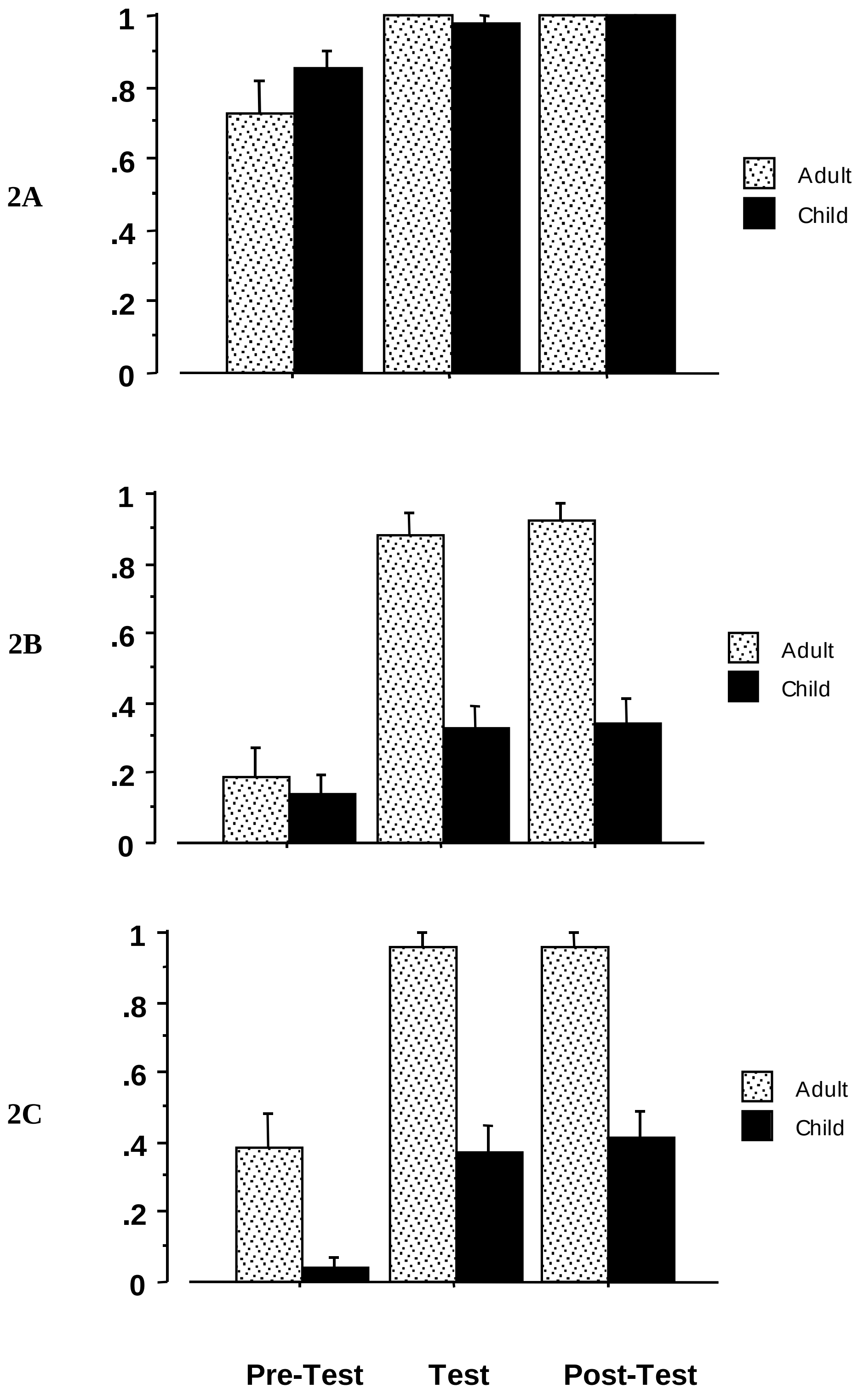
Figure 2 The percent of participants in each age group who believed each variable made a difference at pre-test, test, and post-test phases. Figure 2a shows these results for length of string, Figure 2b for weight of bob, and Figure 2c for release height.

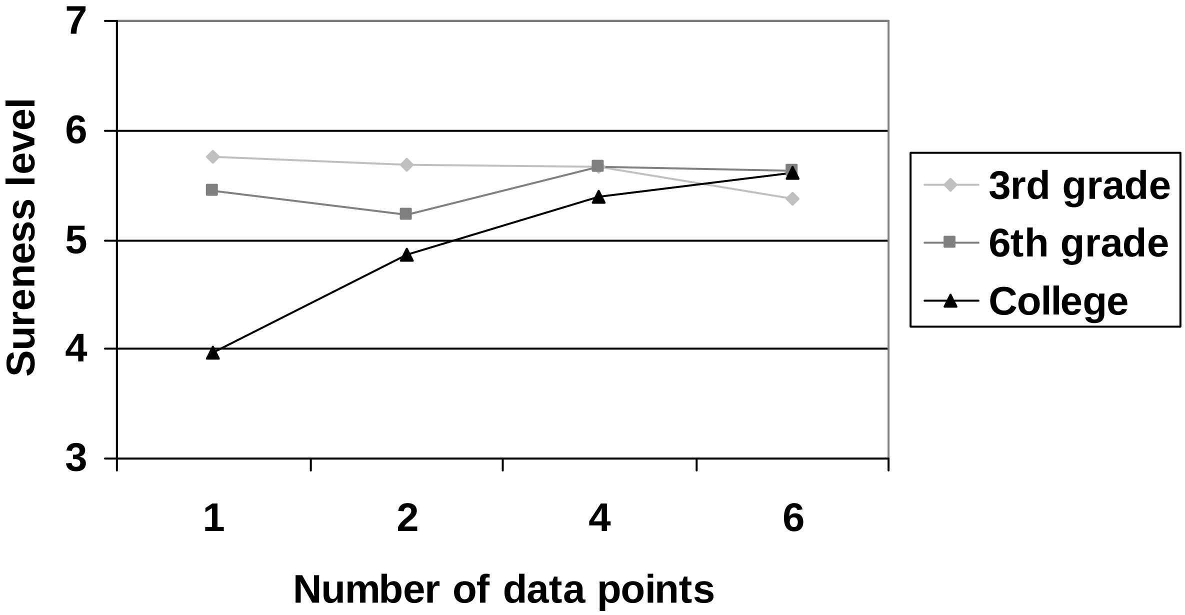
Figure 3 Sureness ratings for different levels of sample size (no overlapping data points), by age.


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


15
14
13
12
11
10
time
heavy
light
Long length, low start


15
14
13
12
11
10
time
high
low
Long length, heavy weight


2A
1
.8
.6
.4
.2
0
Adult
Child
2B
1
.8
.6
.4
.2
0
Adult
Child
2C
1
.8
.6
.4
.2
0
Adult
Child
Pre-Test
Test
Post-Test


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