Elementary School Children's Understanding of Experimental Error

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

Abstract

Understanding data variability and potential error sources is essential to a full understanding of experimental science. We propose a typology of error that considers not only the nature of the error, but also the phase in the experiment in which it occurs. We looked at second- and fourth-grade children's understanding of error, their use of evidence in guiding this understanding, and the role of context in reasoning about error. We found that children could name and recognize sources of error even when they were unable to design controlled experiments. Children used evidence to guide their reasoning, making predictions and drawing conclusions based on the design of their experiments. Children were also sensitive to the context of reasoning: they differentiated the role of random error in relative and absolute measurements. These findings suggest that children understand a wide variety of potential error sources several years before they have acquired the systematic procedures necessary to control and interpret such error

A student in a fourth grade science laboratory is attempting to determine the effect of different factors on how far a ball rolls down a ramp. The ramp is adjustable for length, height, and surface smoothness, and there are two types of balls. Distance is measured by counting the number of discrete "steps" the ball travels up a sloped staircase at the end of the ramp. The teacher's primary instructional goal is to have the child learn how to design unconfounded experiments, in which experimental contrasts differing on only one attribute support logically grounded causal inferences.

In an ideal world, the student could demonstrate an understanding of experimental design by setting up two ramps such that they differed only with respect to surface type, releasing two identical balls, and observing that the ball on the smoother ramp went faster.

But the world is not ideal, and many unanticipated, unknown, and unintended events could influence the outcome of this experiment. Perhaps the student does not fully understand the logic of unconfounded experimentation, and so varies more than one factor at a time. Perhaps one ball is pushed slightly at the start of its roll or hits the side of the ramp on the way down. Perhaps one ball rolls back a few steps at the end of the run and the true distance it rolled is lost. Perhaps none of these "obvious" things occur, but, upon repeating the experiment several times, the student discovers that the balls travel different distances on each replication. What is the student to conclude?

Each of these events can be considered as a different type of experimental error. Among philosophers of science the question of how to best classify different types of experimental error remains controversial. The "conventional" view is that there is a true value and an error term that is part of the observed measurement, and it is distinguishing the magnitude of each that is the difficult part (cf., Hon, 1989). Quite a different perspective – and the one adopted in our analysis – is a process-based view that recognizes the inevitability of errors and that classifies them according to when in the overall cycle of experimental investigation they occur. Hon (1989) has proposed such a taxonomy and used it to organize a wide range of historical cases in which error played an important role in the scientific discovery process. We have adapted his approach by combining it with an earlier, psychologically oriented, classification of error (Toth & Klahr, 1999) to produce a similar taxonomy with which to organize some of the psychological literature on children's understanding of error, and to motivate our own investigations. In our taxonomy, there are five relatively distinct stages to the experimentation process (design, set-up, execution, measurement, and analysis of results), and each stage is associated with a category of error.

Types of Error in Designing and Executing Experiments

Design error This type of error occurs during the earliest conceptual phase of an experiment when some variables not being tested are not controlled and a confounded design is produced. For example, if the goal is to determine the effect of different ramp surfaces, then an experiment that compares a high smooth ramp with a low rough ramp contains a design error. No matter what the outcome of the test, it will be unclear whether any differences in the distance the balls rolled are due to the different steepnesses or different surfaces. Design errors occur "in the head" rather than "in the world" because they result from cognitive failures: either from a failure to fully understand the logic of unconfounded contrasts, or from inadequate domain knowledge.

Execution error These errors occur when something not considered or planned for in the design 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).

Measurement error This type of error overlaps the set-up and the measurement phases, because measurement is involved in setting up the apparatus and calibrating instruments as well as in assessing outcomes.

Interpretation error Interpretation errors can occur at any phase in the experiment. If an error occurs in any of the phases and is not recognized as such, it can influence the interpretation. The analysis phase involves both statistical analysis and theoretical inference, both of which are subject to a wide variety of statistical and cognitive errors. Errors that involve ascribing effects when in fact there are none, or claiming a null effect when one actually exists fall into this category.

Children's Understanding of Experimental Error

Although there is only a small philosophical literature on experimental error, there is an even smaller literature on the psychology of experimental error, i.e., empirical investigations of how people understand and interpret various kinds of experimental error. In this section we briefly summarize what is known about children's understanding of error. The terminology and methods of investigation in these developmental studies are quite varied, making it difficult to compare their results, but we summarize them in terms of the classificatory scheme presented earlier.

Several of the studies of grade school children's error understanding have focused on how children reason about repeated measurements and data variability. Varelas (1997) looked at third and fourth grade children's reasoning about repeated measurements when they carried out experiments in groups. She found that most children expected some variability in measurements, though why they expected this variability was not always clear.

Avoiding an error in the interpretation phase involves assessing when an error is sizeable enough to affect conclusions. Schauble (1996) looked at fifth and sixth graders, and non-college adults. One difficulty many children (and some adults) had was in distinguishing variation due to errors in measuring the results from variation due to true differences between the conditions (i.e., between intended contrasts and measurement phase 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; otherwise, they were more likely to interpret the variability as due to error.

Lubben and Millar (1996) found that some high school children still have considerable difficulty understanding data variability, at least in situations in which they are given the data but are not performing the experiments themselves.

Error is a difficult concept to understand, and it seems likely that there are several levels of understanding (Lubben & Millar, 1996). There is evidence that first and second grade children can recognize a good experiment, even when they cannot yet generate one (Sodian, Zaitchik, & Carey, 1991). This finding suggests that children who are unable to generate error-related reasons for data variability may still have a basic understanding of error and therefore be able to recognize error-based explanations as plausible.

Causal Reasoning and Error Understanding

One way to clarify this literature about error is to reconceptualize error as a subset of the more fundamental topic of causal reasoning. Whether we recognize it as error or not, error is always caused by something. As suggested by the taxonomy proposed here, the nature of that something may be different for each phase of the experimental cycle. From this perspective, before one can reason about error in science experiments, it is necessary to be able to reason about causes (Koslowski & Masnick, in press).

In the current study, we set out to explore several aspects of children's understanding of error: their understanding and recognition of different types of error; consistency of reasoning about experimental design and conclusions, the ability to differentiate between the importance of error when comparing relative and absolute measurements; and the use of theory and evidence in justifications for confidence.

We know that children often have difficulty designing controlled experiments (e.g., Chen & Klahr, 1999). However, it is unclear how much they do understand and whether they can reason consistently based on their incorrect designs. If when children design confounded experiments, they make predictions about the outcome based on variables other than the target variable, it would suggest that although they do not fully understand the goals of the experiment, they use background knowledge to reason correctly based on the experiment they have designed.

Another measure of understanding is children's confidence about conclusions. If they understand the importance of a good design, we would expect them to be more confident of conclusions based on the results of an unconfounded experiment than a confounded one. At a more sophisticated level, if children understand the role of random error, they may still not be completely confident of outcomes after just one or two runs of an unconfounded test. They may consider that there are often uncontrollable factors that can affect a result and, by extension, the conclusions drawn.

The question of when error is important enough to alter conclusions is a difficult one. When looking at simple mechanics problems, the question also depends on the experimental context. If the goal is to determine the exact distance a ball will roll down a ramp under certain conditions, even the slightest unintended intrusion can raise questions about the result. But when the goal is to compare the relative distance a ball rolls given two levels of a particular variable, if the difference is sizeable, error is less important.

We designed a study to address several aspects of children's understanding of error throughout all phases of an experiment, to examine what elementary school children know about different types of error. First, we looked at whether children can design unconfounded experiments, make predictions consistent with their designs, and differentiate the role of error in absolute and relative measurements? Next, we looked at whether children generate alternative reasons for variation in repeated measurements, considering the roles of execution and measurement errors. Finally, we looked at whether children recognize potential sources of error.

Method

Participants Participants were 29 second-grade (mean age = 8.1) and 20 fourth-grade (mean age = 10.1) children from a private elementary school in southwestern Pennsylvania.

Materials Materials included two wooden ramps, each with an adjustable downhill track connected at its lower end to a slightly uphill, "staircase" surface. Children could set three binary variables to configure each ramp: the height (high or low), by using wooden blocks that fit under the ramps; the surface (rough or smooth), by placing inserts on the downhill tracks; and the length of the downhill ramp (long or short), by placing gates at either of two starting positions. Finally, children could choose either a rubber ball or a golf ball to roll down either ramp.

In addition, a laminated copy of a scale for indicating confidence (see below) and a stopwatch were used.

Procedure

Children were interviewed individually. All interviews were videotaped for later coding and analysis. During a brief familiarization phase, children were introduced to the ramp materials and the confidence scale (called a "sureness" scale, with levels of totally sure, pretty sure, kind of sure, and not so sure) and were asked a few questions to ensure that they understood how to use all of the materials.

Part One The purpose of Part One was to determine the extent to which children could design unconfounded experiments with these materials, and to assess their ability to differentiate between absolute and relative measurements. Each child was asked to design four experiments to determine the effect of different settings for specific variables that might affect how far a ball rolls down a ramp. In the first two experiments, children were asked to set up the ramps to test whether the steepness of the ramp made a difference in the outcome; in the third and fourth experiments, they were asked to test the effect of surface.

After the child set up the ramps, the experimenter asked why the ramps had been set up that way and also asked which ball was expected to go farther and why. Next, the experimenter asked the child to release both gates at the same time to see how far the balls rolled.

After the balls had stopped rolling, the experimenter asked the child what he/she had learned and why. Next, the experimenter asked the child whether the target variable (steepness or surface) made a difference. The child used the sureness scale to indicate confidence that the particular variable did make a difference, and to explain why.

Next, the experimenter asked a series of questions about relative and absolute values of the outcome variable. The experimenter asked the child to imagine, if the identical experiment were to be repeated, whether the same ball would go farther, and then whether the two balls would be expected to land on the exact same steps. The children were then asked to rate their sureness about each answer and explain why.

After each experiment and question series, the ramps were disassembled and the child was asked to set up the next experiment.

Part Two The purpose of this part was to explore the child's understanding of data variability in replicated experiments. A single ramp was set up with a high steepness, smooth surface, long run, and a golf ball. For each of five trials, the child was instructed to release the ball by lifting the gate on the experimenter's signal, while the experimenter simultaneously started a stopwatch. When the ball reached the bottom of the ramp (but before it began to roll up the steps), the experimenter stopped the stopwatch and read out a time for the child to record by writing it down. To ensure that all children were presented with the same range of data, the experimenter reported a fixed, predetermined set of times to each child, regardless of the actual time on the stopwatch[1].

At the completion of the five trials, the experimenter noted that it appeared to take a different amount of time for each roll and asked the child to generate reasons to explain these differences. Each child was encouraged to give as many reasons as he or she could think of, and then was asked for a summary explanation he or she would provide the teacher if she asked how long it took.

The experimenter then changed the surface of the ramp to a rough surface, and the ball was again rolled down five times. Again, the experimenter read each child an identical list of run durations. In this second round of numbers, there was a noticeable outlier among the numbers given[2]. After the five trials were completed, the experimenter again asked the child for reasons why the numbers would come out differently and a summary for the teacher.

Part Three Whereas Part Two required children to generate potential sources of error, in Part Three a few such sources were provided, to see how well children

[1] Times: 1.08, 1.20. 1.15, 1.02, 1.17; mean = 1.12, sd = 0.07

[2] Times: 1.90, 2.48, 1.88, 1.95, 1.85; mean = 2.01, sd = 0.26

could reason about their possible influence. Questions about both relative and absolute differences were asked.

The experimenter explained that she had been working with some children at another school who were trying to figure out whether run length made a difference. She demonstrated their situation by presenting the two ramps set up as an unconfounded experiment comparing the short and long run length, with both ramps having high steepness, smooth surfaces, and rubber balls.

The experimenter then asked about three scenarios the students had encountered: 1) one ball hit the side of the ramp on the way down; 2) the two balls were released at different times instead of simultaneously; 3) one ball rolled back a few steps before anyone could record how far it went. Note that scenarios 1 and 3 might be expected to affect both relative and absolute outcomes, while scenario 2 was designed as a control question because it should not effect on the outcome. For each, the experimenter asked the child whether the event described could affect how far the ball went, and whether it could change which ball went farther.

Results

Experimental Design Skills

Children's experimental design skills were assessed by looking at the number of correct (unconfounded) experiments designed. A correct design contrasted the target variable while holding the other three variables constant. This assessed their ability to avoid design phase errors, and to demonstrate knowledge of the Control of Variables Strategy (CVS). We categorized two types of incorrect designs: confounded (contrast of the target factor and one or more other factors), or non-contrastive (no contrast of the target factor). There was a significant effect of grade on CVS performance, with second graders averaging 16% unconfounded experiments, and fourth graders averaging 40% ($t(47) = 2.89$; $p = 0.006$).

Predicting Experimental Outcomes

Children's predictions about which ball would go farther were coded as correct or incorrect. (If the two balls traveled the same number of steps, children were coded as predicting incorrectly, unless they predicted a tie.)

Overall, children were extremely good at predicting the outcomes of unconfounded experiments and significantly less accurate when predicting the outcomes of non-contrastive designs, as measured by Fisher's exact tests of association. [3] For all but the first experiment, there was a strong relationship between predictive accuracy and type of design (unconfounded, confounded, or non-contrastive). (See Table 1.) Children's predictions were most accurate when they designed unconfounded experiments, next most accurate when they designed confounded experiments, and least accurate when they designed non-contrastive experiments. The relationship was stronger for fourth graders than for second graders.

Table 1: Prediction accuracy, by type of experiment

Number Accurate	Uncon-founded	Con-founded	Non - contrastive
Expt. 1	10/10	28/32	5/7
Expt. 2**	9/9	20/27	5/13
Expt. 3*	12/18	16/21	3/10
Expt. 4**	14/14	14/21	6/14

*$p < 0.05$; **$p < 0.01$

Explanations for Predictions

Children's explanations for their predictions were coded for mention of the target variable (steepness or surface), non-target variables, and any prior outcomes.

To assess the consistency of responses, we examined the relationship between children's reasons for their predictions and the type of experiment they designed (unconfounded, confounded, or non-contrastive), using Fisher's exact tests of association. For all four experiments, there was a significant relationship between mention of the target variable and the type of experiment designed. Overall, children mentioned the target variable as an explanation for 92% of their unconfounded experiments, 61% of their confounded experiments and 14% of their non-contrastive experiments. When broken down by grade, there is still a significant relationship at both grade levels. Similarly, there was a significant relationship between type of experiment and mention of the non-target variable. Children said they based their prediction on one or more of the non-target variables 6% of the time when they designed unconfounded experiments, 56% of the time when they designed confounded experiments, and 61% of the time when they designed non-contrastive experiments.

Confidence

Children's responses to the questions, "Can you tell if X makes a difference?" were coded as yes/no responses. Children noted how sure they were of this answer by using the four-level confidence scale. Finally, children explained why they chose the sureness value they did.

Nearly all of the children were "kind of sure," "pretty sure," or "totally sure" about whether steepness makes a difference on the first two experiments (98% and 86%) and whether surface makes a difference on the third and fourth experiments (90% and 86%). Confidence was unrelated to whether the test was unconfounded, as assessed by Fisher's exact tests of association ($p > 0.10$ for each of the four experiments), but there is some evidence that it is related to the accuracy of prediction. Four Fisher's exact tests were performed to assess the relationship between accuracy of prediction and

[3] Six times children designed correct experiments but inaccurately predicted the outcome. These six comparisons all occurred during the third experiment, a comparison of the surfaces, in which either the two balls tied or the ball on the rough surface actually rolled farther than the ball on the smooth surface.

confidence in conclusions. The trend was significant in the third and fourth experiments. In the first experiment, not enough children were unsure to allow for a strong comparison. Children who correctly predicted the outcome were more likely to be sure than those who predicted incorrectly (See Table 2 for details).

Table 2: Percent sure of target variable's effect, by prediction accuracy

Percent sure	Correct prediction	Incorrect prediction	P-value
Expt. 1	98% (42/43)	100% (6/6)	1.000
Expt. 2	91% (31/34)	73% (11/15)	0.179
Expt. 3	97% (30/31)	78% (14/18)	0.054
Expt. 4	94% (32/34)	67% (10/15)	0.022

Comparing Relative/Absolute Replications

For each question about whether the same ball would go farther if the experiment were to be repeated, children first answered yes or no, and then rated their confidence. These two responses – yes/no and confidence level – were combined into a single 7-point ordinal variable, ranging from totally sure the same ball would not go farther to totally sure it would go farther, with not so sure as the midpoint. An analogous coding scheme was used for the questions about whether the balls would land in the exact same positions.

The reasons given for why children expected the same or a different outcome were coded for mention of any of a list of common responses, including the fact that nothing had changed, and the effect of the target variable (e.g., "This one is the steeper ramp so that will make it go farther").

When asked whether the same ball would go farther were the experiment repeated without changes, over 90% of the time children thought it would (i.e., they said that they were kind of sure, pretty sure, or totally sure that it would). This figure excludes cases in which the balls traveled the same distance. The expectations about whether the balls would land in the exact same position were more varied. About 50% of the time, children thought the two balls would not land in the same positions again, about 40% of the time children thought they would, and the remaining times they were unsure.

To test whether children had different expectations for replication of relative and absolute outcomes, scores from the 7-point confidence code for absolute replication were subtracted from the corresponding scores for relative replication. For each child, we computed the mean of this difference score over the four experiments. Children were significantly more confident that the same ball would go farther than that the two balls would land in exactly the same place (mean difference = 2.46; sd = 1.76; t (48) = 9.8; $p < 0.001$). A t-test indicated a marginally significant effect of grade (mean for 2nd grade = 2.1, 4th grade = 3.0, t (47) = 1.95, p = 0.057). Children had different ideas about the importance of variation in the data depending on whether the judgments were about relative or absolute measurements.

Children's reasons for confidence about relative replication were nearly evenly divided: 40% were evidence-based (e.g., "this ball went farther last time"), and 45% theory-based (e.g., "this one will go farther because it's on the steeper ramp").

Accounting for Variability in Replications

Children's explanations for why the timing was different for each of the five trials were coded for mention of several factors, such as the child releasing the gate before or after the experimenter said "go," the experimenter stopping or starting the stopwatch too early or late, or the ball hitting the side of the ramp. Coding agreement over ten participants ranged from 85-100% on each code.

The average number of error sources named on the two sets of trials was 1.48 by second-graders and 2.15 by fourth-graders, a significant difference (t(46) = 2.73; p = 0.009). General linear models examining whether ability to design unconfounded experiments (as assessed by CVS score) is related to ability to name error sources in this second part indicate no relationship once grade is controlled in the model (F (1, 45) = 1. 79, p = 0.19).

Hypothetical Scenarios

Responses to the questions about hypothetical scenarios were classified into one of three categories: (1) yes, with mechanism explanation; (2) yes, without mechanism explanation; (3) no.

All participants correctly said that the ball hitting the side and that the ball rolling back a few steps could influence how far a ball went and whether the same ball would go farther. Eighty-eight percent were able to offer a reason for the former, and 72% were able to offer a reason for the latter. For the question about whether the timing of gate release would affect the distances traveled, 68% said that it would not and 4% offered a plausible mechanism for why it might make a difference (e.g., the vibration of the ramp might be different when a ball is simultaneously rolling down a ramp right next to it).

Overall, 34% of children completely and accurately answered all 6 questions. Fifty percent of the fourth graders and 22% of the second graders answered all the questions correctly.

Generating Error Sources

At several points throughout the interview, children were asked to think of reasons why experiments did not or might not have the same results when repeated (e.g., explaining why the balls would not land in the same place or why a ball rolled down the same ramp five times took a different amount of time for each run). The responses were coded for mention of possible sources of error, such as the ball hitting the side, or wind blowing, or the gates being lifted different ways, as a reason for the variation in results. Eighty-eight percent of children were able to

name at least one source of error. All six children who did not name any sources of error were second-graders.

Discussion

We set out to examine children's knowledge of different error sources. Our results indicate that despite children's difficulty in designing unconfounded experiments, they do understand a lot about error and its importance.

As found in earlier studies (Chen & Klahr, 1999; Toth, et al., 2000), children frequently made design errors (that is., they had difficulty designing unconfounded experiments). However, children showed consistency in their reasoning by referring to their design in justifying their predictions, regardless of whether it was a good experiment.

This evidence of consistency in justifications and conclusions also indicates some causal reasoning abilities. In the majority of cases, children recognized the link between the design and the outcome by considering their design to determine which factors would affect the outcome, and which factors they could draw conclusions about with confidence.

In addition, children's prior theoretical knowledge guided their reasoning. They were more likely to be confident about their conclusions when the evidence matched their prior beliefs (their predictions), though they were still confident more often than not, regardless of their predictions. Children also justified most predictions based on expected effects of the target and non-target variables (though this reliance varied based on design).

Children also demonstrated some understanding of the role of error in interpretation. Even by the second grade, children differentiated the importance of error in different situations, recognizing that errors are much more likely to affect measurements of exact positions than of relative positions. This finding may suggest a nascent understanding of the difference between main effects and specific examples. Children's confidence that the relative ordering would remain the same suggests they expect main effects to be robust, whereas their lack of confidence in absolute outcomes remaining the same suggests their understanding of variability in each sample.

When reasoning about experiments with ramps, children can use several different kinds of information. They can use their domain knowledge, i.e., what they know of the mechanics of friction and gravity, and of other factors that might affect how a specific instrument works. They can also use any formal experimental knowledge they have about what kinds of factors make for a good experiment, such as how to avoid errors in all phases of the experiment. Domain-specific knowledge enables them to name potential sources of error that could affect the outcome, while domain-general knowledge about experimental design encourages them to search for these specific examples.

The fact that most children are able to name at least one possible source of random error indicates that they do have at least a rudimentary idea about how unpredictable and uncontrolled factors can influence an experiment's outcome. At the same time, the observation that most of these same children are not consistently able to design unconfounded experiments suggests that the understanding is not complete at this age. Knowledge about this gap in understanding can lay the foundation for future research about children's knowledge of science experimentation and about the most effective means to aid science educators teaching children these skills.

Acknowledgments

This work was supported in part by grants from NICHHD (HD25211) and the James S. McDonnell Foundation (96-37). We thank Anne Siegel, Jolene Watson, and three anonymous reviewers for comments on earlier drafts.

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.

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

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

Koslowski, B., & Masnick, A. (in press). The development of causal reasoning. In U. Goswami (Ed.) *Blackwell Handbook of Childhood Cognitive Development*. Oxford, England: Blackwell Press.

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

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

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

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

Toth, E. E., & Klahr D. (1999). "It's up to the ball": Children's difficulties in applying valid experimentation strategies in inquiry based science learning environments. Paper presented at the Annual Convention of the American Educational Research Association. Montreal, Canada.

Toth, E. E., Klahr, D., & Chen, Z. (2000). Bridging research and practice: A cognitively-based classroom intervention for teaching experimentation skills to elementary school children. *Cognition and Instruction 18*, 423-459.

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.