JOURNAL OF COGNITION AND DEVELOPMENT, *4*(1), 67–98

Error Matters: An Initial Exploration of Elementary School Children's Understanding of Experimental Error

Amy M. Masnick and David Klahr
Department of Psychology
Carnegie Mellon University

Error is a pervasive and inescapable aspect of empirical science, and it often plays a causal role in experimental outcomes. But little is known about children's understanding of the causes and consequences of experimental error. In this article, we propose a new framework for characterizing experimental error and we use that framework to guide an empirical assessment of elementary school children's understanding of error, their use of theory and evidence in guiding this understanding, and the role of context in reasoning about error. We found that 2nd- and 4th-grade children could both propose and recognize potential sources of error before they could design unconfounded experiments. They used evidence to guide their reasoning, making predictions and drawing conclusions based on the design of their experiments, and they were sensitive to the context of reasoning: They differentiated the role of error in relative and absolute measurements. Long before children have acquired the formal procedures necessary to control error, they have a surprisingly rich—albeit unsystematic—understanding of its various sources.

Error is a pervasive and inescapable aspect of empirical science. Indeed, the acquisition of a deep understanding about different types of error and procedures for dealing with them constitutes an important part of the professional training of scientists in all disciplines. However, the extensive literature on the development of scientific reasoning processes (e.g., Frye, Zelazo, Brooks, & Samuels, 1996; Gopnik, Sobel, Schulz, & Glymour, 2001; Inhelder & Piaget, 1958; Klahr, 2000; Klahr & Simon, 1999; Koslowski, 1996; Kuhn, Amsel, & O'Loughlin, 1988; Schauble, 1996; Siegler, 1975; Siegler & Liebert, 1975) has devoted little attention

Requests for reprints should be sent to Amy M. Masnick, Department of Psychology, Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: masnick@andrew.cmu.edu

to how children come to understand, interpret, and account for error when reasoning in experimental contexts. In this article we propose a new framework for characterizing children's understanding of experimental error, and we use that framework to investigate the way in which second- and fourth-grade children reason about error and its effect on experimental outcomes.

Imagine a child in a fourth-grade science laboratory attempting to determine the effect of different factors on how far a ball travels after rolling down a ramp. Two ramps are provided, both adjustable for length, height, and surface smoothness, and there are two types of balls. A specific goal might be to set up the apparatus to determine how the surface of the ramp (rough or smooth) affects the outcome. In an ideal world, this goal could be accomplished by setting up the two ramps such that they differ only with respect to surface type, releasing two identical balls, and observing how far each travels.

But the world is not ideal, and when children conduct experiments—even very simple ones such as just described—unintended and unanticipated events may influence the outcomes. For example, one ball might be given a slight push at the beginning of its roll or might hit the side of the ramp on the way down. Or perhaps the child neglects to set a variable at the level actually intended, or mistakenly uses a different type of ball on each ramp, or errs in measuring the final outcome. Such intrusions may lead to ambiguous outcomes or inconsistent results, either over time or across "identical" replications by other children in the laboratory. If such intrusions are noticed by an adult supervisor (e.g., a classroom teacher), then the adult might suggest that the child simply rerun the experiment, without providing any explanation for why that trial should be discarded—and forgotten (Toth, Klahr, & Chen, 2000).

However, when such guidance is absent, children must decide (a) how to account for unexpected variability, (b) whether an error has occurred, and (c) to what extent it affects their conclusions. How well can they do this? What do early elementary school children know and understand about experimental error, and how do they integrate such understanding into the design, execution, and interpretation of their experiments? These questions—of relevance to both the psychology of early scientific reasoning and to elementary science education—motivate this investigation.

For children, determining which factors cause which effects can be a daunting task, even in the highly simplified contexts they typically encounter in their earliest school science experiments. Of course, error is not confined to children's experiments, and it plays a role in all empirical research. Consequently, a primary goal of the formal procedures associated with experimental design and data analysis is to minimize the effects of error. Nevertheless, neither the epistemology nor the psychology of experimental error is well understood. Among philosophers of science the question of how to classify different types of experimental error remains controversial (cf. Hon, 1989; Sheynin, 1966). The conventional (statistical) view is that errors are “a tiresome but trivial excrescence on the neat deductive structure of science” (Mellor, 1967, p. 6). On this view, there is a true value and an error term in every measurement, and the difficult part is distinguishing the magnitude of each. Quite a different perspective—and the one adopted in our analysis—is a process-based view that recognizes the inevitability of errors and classifies them according to when they occur in the process of experimental investigation. From this perspective, the statistical definition of error is but one of several types of error that can occur. Hon proposed a taxonomy that classifies errors according to the stage of an experiment in which they occur, and he used it to organize a wide range of cases from the history of scientific discovery in which error played an important role.

In this article, we propose a new taxonomy of experimental error that combines Hon’s (1989) approach with a psychologically oriented classification of error, first proposed by Toth and Klahr (1999). After describing the taxonomy, we use it to structure our review of the small literature on children’s understanding of error. Then we use it as an idealized model of error understanding and organize our investigation in terms of where and how much children’s error understanding deviates from this model.

STAGE-RELATED TYPES OF EXPERIMENTAL ERROR

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 the simple ramps experiment depicted in the opening scenario and used in the investigations described later in this article. However, the taxonomy, shown in Table 1, can be applied to a wide variety of experimental situations.

We distinguish five stages in the experimentation process: design (choosing variables to test), setup (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.

TABLE 1
Experimentation Stages and Error Types

	Stages of Experimentation				
Error Type	*Design (Choose variables to test)*	*Setup (Physically prepare experiment)*	*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

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).[1]

Measurement Error

Measurement error can occur during either the setup stage or the outcome-measurement stage. Error in the setup stage is associated with the readings and settings involved in arranging 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 not considered or anticipated 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).

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

[1] The confounds eagerly detected by reviewers of research proposals and journal manuscripts rarely result from investigators’ lack of knowledge about how to control variables but, instead, from their failure to believe that a particular uncontrolled variable was relevant to the domain they were investigating. As Hon (1989) notes, the history of science is replete with just such errors, many of them having important consequences.

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 effects when in fact there are 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.

PREVIOUS STUDIES OF CHILDREN'S UNDERSTANDING OF EXPERIMENTAL ERROR

Although error is inevitable in all experimental venues—from the elementary school science classroom to world-class research laboratories—little is known about how people understand its sources and ramifications. Moreover, what little psychological research there is has focused mainly on adults (Chinn & Brewer, 1998; Doherty & Tweney, 1988; Freedman, 1992; Gorman, 1986, 1989; O'Connor, Doherty, & Tweney, 1989; Penner & Klahr, 1996). Indeed, we could find only a handful of studies—described later in this article—that explore young children's understanding of different types of error.

In this section we briefly summarize what is known about children's understanding of error, and then we propose several questions to be addressed in this study. The relations among type errors at different stages of experimentation have not been elucidated in the studies reviewed here because most of them focus on just one of the five stages. Moreover, the terminology and methodology in these studies are quite varied, making it difficult to compare results. Here we summarize them in terms of the classification scheme presented earlier.

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 noncollege 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 (in press) 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.

Lubben and Millar (1996) investigated children's understanding of execution error and measurement error. Children age 11 to 15 (American Grades 4–9) were asked to make judgments about the importance of repeating measurements and about the interpretations of data variability. They found that some high school students 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.

Taken as a whole, this small collection of studies does not lead to any robust conclusions about children's error understanding. 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.

CAUSAL REASONING AND ERROR UNDERSTANDING

> One of the problems with science is that experiments do not always turn out the way that we expect. Sometimes data is obtained that is due to error; other times, scientists interpret data erroneously—mistakenly assuming that a particular cause generated a certain effect. … [Hence] … much of dealing with error is a way of deciding what is the cause of a particular effect. Dealing with error is therefore a type of causal reasoning that scientists must constantly grapple with. (Dunbar, 2001, p. 129)

Dunbar's observation suggests that it might be possible to clarify the literature about error by situating it within the more extensive literature on the development of causal reasoning. One type of causal reasoning is abduction: reasoning from an existing situation to a hypothesized cause of that situation. Abduction plays an important role in the attempt to identify experimental error because, for any given experimental outcome, one must propose causes that led to the result, determining whether different types of error are likely as possible causes. From this perspective, for children to reason effectively about error in scientific experiments requires that they have fairly strong causal reasoning skills.

Unfortunately, here too, we find contradictory evidence about children's causal reasoning skills. On the one hand, some studies have suggested many flaws in children's causal reasoning, which causes them difficulty in reasoning about science (e.g., Inhelder & Piaget, 1958; Kuhn et al., 1988; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995). On the other hand, several other studies and reviews have suggested that children do have a solid grasp of causality from an early age (e.g., Bullock, Gelman, & Baillargeon, 1982; Gopnik & Sobel, 2000; Gopnik et al., 2001; Koslowski & Masnick, 2002; Shultz, 1982; Siegler, 1975), perhaps even as young as 10 months old (Oakes & Cohen, 1990). The varying methodologies of these studies may partly account for the different outcomes: Studies with explicit measures tend to indicate that children have some difficulties, whereas studies with implicit measures suggest that they are more knowledgeable.

However, most of the research on causal reasoning ignores the issue of error (e.g., Amsel & Brock, 1996; Bullock et al., 1982; Shaklee & Paszek, 1985). A common practice is to present children with covariation matrices in which the frequency of various co-occurrences is presented (e.g., number of instances of healthy or sick plants with or without fertilizer). Children are asked to make causal inference from these types of data distributions, but the question of why there are any instances in the off-diagonal cells is not addressed. That is, why were only some, rather than all, of the fertilized plants healthy? Even when a mechanism is provided for children to reason about causation (e.g., Frye et al., 1996; Schlottman, 1999), it is typically a discrete task that does not have variation and therefore does not examine the role of error in general.

Children's understanding of causation in general plays a role throughout experimentation. At all stages in an experiment, the same set of causal factors must be considered. Some causes are anticipated as effects of the experimental design. Other potential causes are considered errors because they interfere with what the experiment is designed to measure and can affect the conclusions that can be drawn from the results. At the start of an experiment, the goal is to minimize the influence of such errors; at the end of an experiment, the goal is to determine which errors did occur and the extent to which they affect the conclusions that can be drawn. If children know that the outcome is determined by all of the things that happen during the experiment (i.e., that it is causally determined), then they can apply this knowledge at each stage of the experiment. Thus, children may be able to draw on this casual knowledge when choosing the experimental setup, reasoning about factors that could affect the execution and measurement, deciding what conclusions to draw from the results, and deciding how much confidence to place in these results. Consequently, both causal reasoning skills and beliefs about which factors might cause an effect (and what kind of an effect they might cause) play critical roles in each stage of the experiment.

THIS STUDY

Previous investigations of children's error understanding, and our consideration of the role of causal reasoning in this understanding, still leave many questions unanswered. Without explicit instruction, most third and fourth graders tend not to design unconfounded experiments. But they can be taught these skills quite easily (Chen & Klahr, 1999; Toth et al., 2000). This finding suggests that these children have the beginnings of an understanding about experimentation; without this foundation, they would be unlikely to learn the skills so quickly after brief instruction. Although it is possible that extensive knowledge about the different kinds of error precedes a full understanding of how to design an unconfounded experiment, the subtleties of several types of error make that conjecture improbable. More likely, children's ideas about error and experimentation develop concurrently. Thus, children in the early elementary school grades are an appropriate population for investigating the emergence of error understanding.

We know that uninstructed fourth graders often have difficulty in designing controlled experiments (Chen & Klahr, 1999), but can they use the design they created to predict outcomes correctly? Sometimes, they may correctly predict the outcome of a confounded experiment if the confounded factors all happen to point in the same direction. For example, a confounded comparison between a high ramp with a smooth surface and a low ramp with a rough surface would still allow a child to correctly predict that the ball on the higher ramp will roll farther. Correct predictions in such situations would suggest that although children do not fully understand how to isolate the causal role of a single factor, they are able to base their prediction on domain-specific knowledge: that is, they can use their knowledge about ramps, rather than the logic of unconfounded experimental design. This level of performance demonstrates some understanding of the connection between design and outcome even when an understanding of how to design an appropriate test of a hypothesis is not yet present.

Children's confidence in their conclusions provides another measure of their understanding of how the different stages of an experiment collectively affect the outcome. If they understand the importance of a good design, then they should be more confident about the role of a specific factor when their conclusion is based on the results of an unconfounded experiment than when it is based on a confounded one. Children with a more sophisticated understanding of the unpredictable nature of execution and measurement errors may still not be highly confident about the results of only a few runs of an unconfounded test. They may recognize and consider that there are often uncontrollable factors that can affect a result and, by extension, the conclusions drawn.

Similarly, children's understanding of how repeated measurements often yield different results can be indicative of some causal understanding of error. There is some recent evidence that children use several characteristics of data, including their theoretical understanding about variation, to make inferences about data sets (Jacobs & Narloch, 2001).

The question of when error is important enough to alter conclusions is, in effect, a question about statistical significance and effect size, topics not usually taught in detail until the college level. Even adults have difficulty with the concept (e.g., Schauble, 1996). However, when looking at simple mechanics problems, the question also depends on the exact experimental context. The precision required is intrinsically related to the goal of the experiment. 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 sizable, error is less important. Therefore, a key part of understanding error and variability in science is knowing when error matters.

When children reason about experiments and error, they can draw on knowledge about the content domain or knowledge about experimentation. Domain-specific knowledge could include such things as what they know of the mechanics of friction and gravity and of other factors that might affect how a specific instrument works. Domain-general knowledge could include an understanding about what kinds of factors make for a good experiment, such as that all known factors should be controlled, that it is important to consider any source of variation in the execution, and that the variables and outcome must all be measured as accurately as possible. Domain-specific knowledge enables children to name potential sources of error that could affect the outcome, whereas domain-general knowledge about experimental design encourages them to search for specific examples.

The study described here was designed to address several questions about children's error knowledge throughout all stages of an experiment. First, in the design stage, can children create unconfounded experiments and make predictions consistent with their designs? Second, can children differentiate the role of error in absolute and relative measurements? Third, are children able to generate alternative reasons for variation in repeated measurements, and are they able to consider the role of different sources and consequences of error in the different experimental stages? Fourth, can children recognize potential sources of error? Fifth, what are the relations among children's understanding of the different types of errors?

We chose to examine this topic in the domain of ramps. Although ramps are simple mechanical structures, there are several points at which different factors can influence the outcome, leading to some variation in results. In addition, ramps are a familiar domain about which children are likely to have some causal knowledge to draw on in reaching conclusions.

METHOD AND RESULTS

General Method

Participants Participants were 29 second-grade (mean age = 8.1 years; range = 7.4–9.2) and 20 fourth-grade (mean age = 10.1 years; range = 9.5-10.7) children from a private elementary school in southwestern Pennsylvania. The children were recruited from letters sent to parents.

Materials Materials included two wooden ramps, each with an adjustable downhill track connected at its lower end to a slightly uphill "staircase" surface. (See Figure 1.) 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 a ball to roll down each ramp.

There are 18 steps on the uphill ramp, with a run of 1 in. (2.54 cm), and a rise of ¼ in. (.635 cm) *Steepness:* The ramp was set at either 8° or 11° of steepness. Across all variations in setup, the mean difference in the effect size of steepness is 4.7 steps. *Surface:* The ramp insert was placed either with a smooth wooden sur-

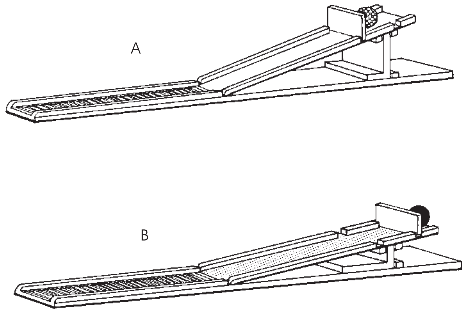

FIGURE 1. The Ramps Domain. On each of the two ramps, children can vary the angle of the downhill slope, the surface of the ramp, the length of the ramp, and the type of ball. The confounded experiment depicted here contrasts (a) the golf ball on the steep, smooth ramp with a short run with (b) the rubber ball on a shallow, rough ramp with a long run.

face, face-up, or with a carpeted surface, face-up. Across all variations in setup, the mean difference in the effect size of surface is 1.8 steps. *Run length:* With a long run, the ball travels 21 in. (53.34 cm) before reaching the sloped steps; with a short run the ball travels 16 in. (40.64 cm) before reaching the steps. Across all variations in setup, the mean difference in the effect size of run length is 1.9 steps. *Ball type:* Standard-size golf balls and rubber squash balls were used; both were approximately 1.5 in. (3.81 cm) in diameter. Across all variations in setup, the mean difference in the effect size of ball type is 0.8 steps.

To set up an experiment, children constructed two ramps, setting the steepness, surface, and length of run for each and then placing one ball behind the gate on each ramp. To run the experiment, children removed the gates and observed as each ball rolled down its ramp and then up the steps until coming to a stop. The dependent measure was the number of steps that the ball traveled up the stepped side of the ramp (the numbers were written on the ramp, next to each step).

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

Procedure. Children were interviewed individually. All interviews were videotaped for later coding and analysis. A 5-min familiarization was followed by a three-phase interview, with each phase lasting 10 to 20 min (total time approximately 45 min per child). For the sake of clarity, we present the results of each phase before describing the next phase.

During the familiarization, the ramp materials were presented to the child, and the experimenter made sure that the child could identify the different values of steepness, surface, and run length. In addition, a 4-level confidence scale (*not so* *sure, kind of sure, pretty sure, totally sure*) was presented and described. To ensure that children understood what the scale was supposed to measure, they were asked a few example questions. All children appeared to understand the scale after two examples.

PHASE I: EXPERIMENTAL DESIGN AND OUTCOME PREDICTION

Phase I Method

The purpose of this phase 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. This allowed us to compare these participants' understanding of the Control of Variables Strategy (CVS) with earlier studies using the same materials (Chen & Klahr, 1999; Toth et al., 2000) and to examine the relation between their CVS scores (indicating design-stage error) and other types of error understanding assessed in later parts. Procedurally, CVS is a method for creating experiments in which a single contrast is made between experimental conditions. Conceptually, CVS involves making appropriate inferences from the outcomes of unconfounded experiments as well as understanding the indeterminacy of confounded experiments. Theoretically, CVS imposes powerful constraints on the size of the space of all possible experiments (Klahr, 2000; Klahr & Simon, 1999).

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 and second experiments, each child was 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, each child was asked to set up the ramps to test whether the surface of the ramp made a difference.[2] After the child set up the ramps, but before the balls were released, the experimenter asked why the ramps had been set up that way. The experimenter also asked which ball the child 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 where each ball had landed, and the child read off the number of the step each ball landed on. The experimenter then asked the child what he or she had learned and why. Next, the ex-

[2] Note that we did not counterbalance focal variable (so our own study confounds experience with our focal variable). However, in several of our previous studies with the same materials (e.g., Chen & Klahr, 1999), the order of focal variables was carefully counterbalanced and never affected the outcome, either as a main effect or an interaction with other independent variables.

perimenter asked the child whether the target variable (steepness for the first two experiments, and surface for the third and fourth experiments) made a difference. The child was then asked to use the confidence scale to indicate how sure he or she was that the particular variable did make a difference, and then to explain why.

Next, the experimenter asked a series of questions about relative and absolute values of the outcome variable. First, for the relative values, the experimenter asked the child to imagine what would happen if the identical experiment were to be repeated. The experimenter asked, “If you put the balls back at the top and don’t change anything, and then let them go again, do you think this one [pointing to ball that went farther] would go farther than this one [pointing to other ball]?”; “How sure are you?”; and “Why?” Next, for the absolute values, similar questions were asked about whether the child expected the two balls to land on exactly the same steps, were the experiment to be repeated.

This sequence of questions in Part 1 was repeated for the remaining three experiments. After each experiment and question series, the ramps were disassembled and the child was asked to set up the next experiment.

Phase I Results

Experimental design skills. Children’s experimental design skills were assessed by scoring their ability to design unconfounded experiments using the two ramps. A correct design contrasted the target variable but held the other three variables constant. There were two types of design errors: confounded design (contrast of the target variable and one or more other variables) or noncontrastive design (no contrast of the target variable). For some analyses, the two incorrect responses were collapsed, to yield a dichotomous variable measuring correct or incorrect comparisons. There was a significant grade difference in frequency of design errors, with second graders averaging 16% unconfounded experiments, and fourth graders averaging 40%, [$t(47) = 2.89$, $p = .006$].[3]

Predicting experimental outcomes. Recall that children designed four experiments. They were asked to predict the relative outcome, that is, which ball would go farther. These predictions could be based on both the design of the experiment and prior knowledge. The predictions were each coded as accurate or inaccurate, based on the actual difference in the number of steps the balls traveled on each ramp.[4]

[3] These numbers are comparable to the findings of Chen and Klahr (1999) in which, prior to instruction, second graders averaged 26% unconfounded experiments and fourth graders averaged 48% unconfounded experiments. Similarly, Toth et al. (2000) found that fourth graders averaged 30% accuracy in designing unconfounded experiments before instruction.

Overall, children were extremely accurate at predicting the outcomes of their unconfounded experiments and significantly less accurate at predicting the outcomes of their noncontrastive designs.[5] Fisher exact probability tests of association were used to assess relations between the type of design and the accuracy of the prediction. There was a consistent relation between predictive accuracy and type of design (unconfounded, confounded, or noncontrastive) for both second and fourth graders, although it reached statistical significance only for fourth graders (see Table 2). Children were accurate 89% of the time when predicting the outcomes of unconfounded experiments but only accurate 43% of the time when predicting the outcomes of noncontrastive experiments. They were accurate 77% of the time when predicting the outcomes of confounded experiments. As noted earlier, a confound does not preclude a good prediction if the values of both of the confounded variables point in the same direction. Indeed, this type of design corresponds to Schauble, Klopfer, and Raghavan's (1991) "engineering" approach in which the child is more interested in obtaining a desired outcome than in isolating its causes. The consistent pattern linking prediction accuracy and type of design suggests a relation between prior domain knowledge and understanding the importance of avoiding design error.

Explanations for predictions. To assess the extent to which children were giving consistent responses, we examined the relation between children's reasons for their predictions and the type of experiment they designed (unconfounded, confounded, or noncontrastive). Reasons were coded for mention of any of the following items: (a) the target variable, (b) any of the nontarget variables, and (c) an outcome from one of the earlier experiments. Two coders independently coded the same 20% of the data. Coding agreement on each item ranged from 90% to 100%. All differences were resolved through discussion, and one coder then coded the remainder of the data.

For each of the four experiments, there was a significant relation between mention of the target variable and the type of experiment, χ^2 (2, $N = 49$) = 22.20, 16.45, 15.29, 14.55 for Experiments 1 to 4, respectively; all ps < .001. Overall, children mentioned the target variable as justification for prediction for 92% of their unconfounded experiments, compared with 61% of their confounded experiments and only 14% of their noncontrastive experiments. The relation holds when each grade is analyzed separately, using Fisher exact probability tests to account for the

[4] If there was no difference in the number of steps traveled by each ball and children had predicted a difference, they were coded as predicting inaccurately. This occurred 14 times out of 196 experiments.

[5] Six times children designed correct, unconfounded experiments but inaccurately predicted the outcome. These six comparisons were all 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.

TABLE 2
Proportion of Correct Predictions for Each Type of Experimental Design

Children's Experiments	*Grade*	*Unconfounded*	*Confounded*	*Noncontrastive*
1				
(Steepness)	2	100% (5/5)	86% (18/21)	67% (2/3)
	4	100% (5/5)	91% (10/11)	75% (3/4)
2				
(Steepness)	2	100% (5/5)	70% (14/20)	50% (2/4)
	4**	100% (4/4)	85% (6/7)	33% (3/9)
3				
(Surface)	2	60% (3/5)	82% (14/17)	43% (3/7)
	4*	69% (9/13)	50% (2/4)	0% (0/3)
4				
(Surface)	2	100% (4/4)	63% (10/16)	44% (4/9)
	4**	100% (10/10)	80% (4/5)	40% (2/5)
All four experiments	2	89% (17/19)	76% (56/74)	48% (11/23)
	4	88% (28/32)	81% (22/27)	38% (8/21)

* $p < .10$. ** $p < .05$.

small sample sizes (all *p*s < .1). Similarly, there was a significant relation between type of experiment and mention of nontarget variables, χ^2 (2, $N = 49$) = 11.65, $p = .003$; χ^2 (2, $N = 49$) = 8.72, $p = .013$; χ^2 (2, $N = 49$) = 10.02, $p = .007$; χ^2 (2, $N = 49$) = 12.00, $p = .002$ for Experiments 1–4, respectively. Children said they based their prediction on one or more of the nontarget variables only 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 noncontrastive experiments. There was no relation between mention of prior outcome and type of design—prior outcome was rarely mentioned as a reason (mentioned a total of 13 times in 196 opportunities).

To summarize, when children designed unconfounded experiments, they almost always based their predictions on the expected effect of the target variables rather than on any of the nontarget variables. In contrast, when children designed noncontrastive experiments, they rarely based their predictions on the target variable. These findings suggest that even when children do not understand how to design an unconfounded experiment, they use the information about the setup they have designed to make predictions.

Confidence in conclusions. Children were asked three questions about their confidence in the conclusions for each experiment. Children's responses to the first question, "Can you tell if *X* makes a difference?" were coded simply as yes/no responses. The second question assessed children's confidence in their response to the first question.[6] Children used the 4-point confidence scale to indicate their confidence (*not so sure, kind of sure, pretty sure, totally sure*). Finally, children were asked why they chose the confidence value they did.

Children's responses to this third question—why they were or were not sure about the effect—were coded for mention of any of the following items: (a) design-based explanations (e.g., "All the other things were the same"), (b) a reference to the current outcome, (c) a reference to any previous outcomes, (d) a belief about the target variable (e.g., "The bumpy surface slows it down"), (e) a belief about one or more nontarget variables, (f) some kind of error, and (g) any other reason offered. Coding agreement on individual items ranged from 70% to 100%.

Nearly all of the children were "kind of sure," "pretty sure," or "totally sure" about whether steepness makes a difference on Experiment 1 (98%) and Experiment 2 (86%). Most of the children were also sure about whether surface makes a difference on the third and fourth experiments (90% and 86%, respectively). Confidence was unrelated to whether the test was unconfounded, but there is some evidence that it was related to the accuracy of prediction. Four Fisher exact probability tests were performed to assess the relation between accuracy of prediction and confidence in conclusions, one for each experiment the children designed. The relation was significant in the third and fourth experiments. In the first experiment, only 1 child was unsure about the conclusions, and in the second experiment, the trend was in the expected direction, even though it did not reach significance. Children who correctly predicted the outcome were more likely to be sure than those who predicted incorrectly (see Table 3). In addition, this trend holds when the confidence data are split into "totally sure" and "pretty sure" compared with "kind of sure" and "not so sure," as well as when the data are split to compare "totally sure" with all other categories. In other words, children who were accurate in their predictions were more likely to be highly confident about their conclusions than those who made inaccurate predictions.

What kinds of reasons did children give for their confidence responses? Although these reasons varied widely, children rarely mentioned potential error sources. Mention of design or execution error was very infrequent (5% and 3% of responses, respectively). No children mentioned measurement error as a reason for either their confidence or lack of confidence in drawing conclusions.

Rather than suggest error sources as a basis for confidence judgments, children tended to use either prior domain knowledge or current empirical evidence (or both) to justify confidence levels. An average of 55% of the time participants offered reasons based on prior domain knowledge (e.g., "I'm sure because steeper

[6] This question was asked in two forms. In one form, children were asked, "How sure are you that you can tell that *X* makes a difference?" In another form, children were asked, "How sure are you that *X* makes a difference?" Subsequent analyses showed no differences in response to the two forms, so the responses were collapsed for analysis.

TABLE 3
Percent Who Were Sure, by Prediction Accuracy, for Each Experiment

Children's Experiments	*Sure, Given Correct Prediction*	*Sure, Given Incorrect Prediction*	*P*
1	98% (42/43)	100% (6/6)	1.000
2	91% (31/34)	73% (11/15)	.179
3	97% (30/31)	78% (14/18)	.054
4	94% (32/34)	67% (10/15)	.022
All four experiments	95% (135/142)	76% (41/54)	

ramps make balls go farther"). In addition, an average of 31% of the time participants offered reasons based on the evidence from the current or previous experiments with the ramps.

Replication of relative–absolute outcomes. 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 on the 4-point scale. These two responses—yes/no and confidence level—were combined into a single 7-point ordinal variable: *totally sure the same ball would not go farther, pretty sure it would not go farther, kind of sure it would not go farther, not so sure, kind of sure it would go farther, pretty sure it would go farther, totally sure it would go farther.* The same coding scheme was used for the questions about whether the balls would come to rest in exactly the same positions.

The reasons given for why children expected the same or a different outcome were coded for mention of any of the following responses: (a) using evidence from the experiment just run or indicating nothing about the setup had been changed, (b) evidence from previous experiments, (c) the magnitude of the difference last time, (d) the effect of the target variable (e.g., "This one is the steeper ramp so that will make it go farther"), (e) the effect of one or more nontarget variables, (f) the ball could or did hit the side of the ramp on the way down, (g) the balls could be or were released at different times from the gates, (h) the way the gate was released might be different, (i) the wind could blow the ball sometimes, (j) some other error might affect the outcome, and (k) some other nonerror factor might affect the results. Items (f) to (j) were considered error-based responses. Coding agreement over 10 participants was at least 95% on each item for the reasons the same ball might or might not go farther, and it was at least 90% on each item for the reasons the balls might or might not come to rest in the exact same positions.

When asked whether the same ball would go farther if the experiment were to be repeated without changes, an average of 94% of the time children thought it would; that is, they said that they were kind of sure, pretty sure, or totally sure that it would (when those who were only kind of sure are eliminated, this number drops to 84%). 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 (that is, they were kind of, pretty, or totally sure that they would not); about 40% of the time children thought they would (that is, they were kind of, pretty, or totally sure that they would) and the remaining times they were unsure. (Considering only cases in which children said they were pretty sure or totally sure indicate that 43% of the time they were confident the balls would not land in the same place, whereas 26% of the time they were confident they would.) However, there were significant grade differences; four Fisher exact probability tests each indicated a relation between grade and confidence that the balls would not land in the same position (*p*s = .026, .009, .023, and .098, for Experiments 1–4, respectively). An average of 74% of the fourth graders were sure the balls would not land in the same positions, whereas only 38% of the second graders were sure they would not land in the same positions.

To test whether children had different expectations for replication of relative and absolute outcomes, scores from the 7-point sureness scale 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 sure about the replication of relative positions than they were about the replication of exact positions with mean difference = 2.46 ($SD = 1.76$), significantly different from zero, $t(48) = 9.8, p < .001$. There was a marginally significant effect of grade [mean for second grade = 2.1, mean for fourth grade = 3.0, $t(47) = 1.95$, $p = .057$]. Thus, there is evidence that children considered whether they wanted to know the relative or absolute distances when considering the importance of variation in the data.

One way to assess children's understanding of effect size is to look at the relation between the actual distance the balls traveled and how confident children are that the same ball will go farther. Four regressions were performed to examine this link. Instances in which the two balls went the same distance were excluded from these analyses. For the first and third experiments, the results were highly significant, whereas for the second and fourth, they were marginally significantly related. For Experiments 1 to 4, the regressions yielded the following values, $F(1, 47) = 13.98, p < .001$; $F(1, 40) = 3.175, p = .082$; $F(1, 40) = 17.66, p < .001$; $F(1, 45) = 3.056$; $p = .087$. Note that the same pattern held for questions about both steepness and surface: When the balls were farther apart, children were more confident that the relative positions would be the same if the experiment were repeated.

Another indication of children's ability to differentiate between absolute and relative measurements comes from differences in the kinds of justifications given for the two situations. Justifications mentioning the ball hitting the side of the ramp, releasing the two balls at different times, releasing the gates differently, wind blowing, or any other outside, unpredictable factor that might affect the results were considered reasons based on execution error. Very few children gave any execution-error reasons for why the same ball might or might not go farther (across the four experiments, an average of 8% of the children mentioned execution error in response to this question). However, children were much more likely to give execution-error reasons for why the balls might not land on the same exact spots (across the four experiments, an average of 37% of the children mentioned error). There were grade differences in the percentage of children offering such reasons for the absolute measurements, with fourth graders generally more likely to mention execution error sources, χ^2 (1, N = 49) = 3.29, p = .070; χ^2 (1, N = 49) = 4.85, p = .028; χ^2 (1, N = 49) = 3.75, p = .053; χ^2 (1, N = 49) = 6.77, p = .009, for Experiments 1 to 4, respectively.

Although children did not often justify their expectations of relative position by referring to execution error, they did offer other justifications. To examine more closely the different sources of information children used in drawing conclusions, children's justifications, which were coded as described previously, were also grouped as evidence based and domain-theory based. When children referred to either the results of this experiment or the results of previous experiments, their justifications were considered to be evidence based. When children justified their answer with a reference to a belief about the effect of one or more of the variables, their answers were considered to be domain-theory based. Grouping the answers this way allows a broader picture of the information children are using when they are not talking about error. Children's reasons for confidence about relative replication were nearly evenly divided: 40% were evidence based and 45% domain-theory based.

PHASE II: DATA VARIABILITY

Phase II Method

The purpose of this phase was to explore children's understanding of data variability in replicated experiments. To control variability and increase precision in this part of our investigation, we changed the dependent variable of interest from discrete distance (on the stepped receiving ramp) to time. 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. (The series of times was 1.08 s, 1.20

sec, 1.15 s, 1.02 s, and 1.17 s; *M* = 1.12, *SD* = 0.07. Children were only read the times, one after each run, and were not provided with the mean and standard deviation.) All children accepted the times given as valid. At the completion of the five trials, the experimenter commented that the child had rolled the same ball down the same ramp five times, and yet it appeared to take a different amount of time for each roll, as one could see by looking at the list of times the child had recorded and could refer to. The child was asked to generate reasons to explain these differences: "Can you think of some reasons why the results came out differently even though we rolled the same ball down the same ramp five times?" The experimenter prompted children who said they could not think of anything: "Let's think about this for a minute. We rolled the same ball down the exact same ramp five times, and yet we got five different numbers. That's a little strange, isn't it? Can you think of any reasons why that might have happened?" Each child gave as many reasons as he or she could think of.

Next, the child's understanding of the statistical idea of "most representative number" was assessed. With the numbers for the five trials still in view, the experimenter asked, "If your teacher asked how long the ball takes to go down this ramp, what would you say?" The experimenter also asked the child why the proposed answer was a good one.

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 the same predetermined list of run durations. In this second round of numbers, there was a noticeable outlier among the numbers given (times were 1.90 s, 2.48 sec, 1.88 s, 1.95 s, and 1.85 s; *M* = 2.01, *SD* = 0.26). After the five trials were completed, the experimenter again asked the child for reasons why the numbers would come out differently and to provide a representative time for the teacher.

Phase II Results

Accounting for variability in replications. Children's explanations for the variation in data were coded for mention of several error types and grouped into two categories for later analyses. The first category consisted of factors that occurred in the measurement of the experiment (either in the setup or the measurement of the outcome). This category included mention of the child lifting the gate before or after the experimenter said "go," and of the experimenter stopping or starting the stopwatch early or late. The second category of possible factors included execution errors. In this category was mention of (a) the particular way the gate was released (e.g., sometimes it was lifted quickly and other times slowly), (b) the initial position of the ball relative to the gate (that is, on the side of the ramp or in the center), (c) the tilt of the gate (set forward sometimes and backward other times), (d) the ball hitting the side of the ramp on some trials, (e) wind blowing the ball, (f) irregularities in the surface of the ramp (e.g., the rough surface may be bumpier in some parts, and the ball may have rolled down different parts of the ramp), (g) other errors in executing the experiment, (h) other specific errors not included on the list (e.g., “One time the table was bumped”), and (i) other vague errors (e.g., “The ball just sometimes goes slower or faster”). Coding agreement over 10 participants ranged from 85% to 100% on each code within both categories.

Children gave several reasons for nonidentical times on identical replications, and there were grade differences. The mean number of different error sources named on the two sets of trials was 1.48 by second graders and 2.15 by fourth graders, $t(46) = 2.73$, $p = .009$. General linear models examining whether ability to name error sources is related to ability to avoid design errors indicated no relation once grade is controlled in the model [for CVS score, $F(1, 45) = 1.79$, $p = .19$].

In addition to the mean differences, there were individual differences in the frequency with which types of errors were mentioned. Measurement errors were named by 47% of children. There was a significant grade difference: 65% of fourth graders named at least one source of measurement error, and 34% of second graders named at least one source, χ^2 (1, $N = 49$) = 4.43, $p = .035$. Mention of execution errors was more common: 88% of children named at least one type of execution error. Here, too, there were significant grade differences: 100% of fourth graders named at least one source of execution error, and 79% of second graders did so χ^2 (1, $N = 49$) = 4.72, $p = .030$.

Choosing the most representative number for variable outcomes. Children’s answers to what they would tell their teacher if asked how long it takes a ball to go down the ramp were classified into one of the following mutually exclusive categories: (a) arithmetic mean, (b) fastest time recorded, (c) one of the recorded times but not the slowest or fastest, (d) a time generated by the child in between the fastest and the slowest times, (e) slowest time recorded, (f) all five times, (g) range of times, (h) other child-generated time (i) don’t know, and (j) other.

The justifications for the appropriate times to tell the teacher were classified into one of the following categories: (a) the single time was from the “best” run; (b) it is a number in the middle of the range of times (either one of the data points read by the experimenter and recorded by the child, or a different time generated by the child); (c) it is the most informative; and (d) other.

There was a wide range of answers for the summary variable, and some categories were merged for analyses. No child suggested using the mean. For the first set of runs, 31% chose a time between the minimum and maximum times (either experimenter-generated or child-generated), 8% of children chose the fastest time, 19% chose the slowest time, 13% chose either all five times or the range of times, 12% chose another time not in the range, and the remaining 17% did not know or gave no clear answer. For the second set of runs, a similar pattern emerged. Thirty-five percent of children chose a nonextreme time, 13% chose the fastest time, 12% chose the slowest time (the outlier), 17% chose either all 5 times or a range of times, 10% chose a time not in the range, and the remaining 12% did not offer a clear answer. Overall, there was no evidence of a grade effect on the distribution of answers. Fisher exact probability tests were used to explore the link between grade level and response patterns because of the large number of cells relative to the number of participants (p = .62 and .40, for the first and second set of runs, respectively).

Arguably, a time in the middle of the range is the best answer. There were no grade differences in selecting this answer on the first set of runs, χ^2 (1, N = 49) = 0.006, p = .938, and a marginally significant relation for the second set of runs, χ^2 (1, N = 49)= 3.49, p = .062, with fourth graders more likely to choose a time in the middle.

Most of the children had difficulty justifying their choice of summary variables. A few children said that they chose their answer because it was the "best run" (4% and 6%, on the first and second sets of runs, respectively), and some said they chose a time in the middle (17% and 21%, on the first and second set of runs, respectively). All of those who said they chose their time because it was in the middle of the numbers did actually choose a time in the middle. All other children either gave no answer, or gave an answer that was not clear.

PHASE III: ERROR SOURCE REASONING

Phase III Method

Whereas the data variability phase required children to generate potential sources of error in this domain, in this phase a few such sources were provided to see how well children could reason about their possible influence. Children were asked about both relative and absolute measurements.

The general procedure was to ask children about hypothetical sources of error (albeit not in those terms) and their effect on the outcome. 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 said that children had been working in groups in their classroom, and she demonstrated their ramp setup 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. For each, the experimenter asked the child whether or not the event described could affect how far the ball went, and whether or not it could change which of the two balls went farther. Children were also prompted to explain their answers.

1. What would happen if one of the balls hit the side of the ramp and the other did not?

2. What would happen if the balls were released at different times, instead of simultaneously?
3. What would happen if one ball rolled back a few steps before anyone got a chance to record how far it went?

Scenario 1 addressed execution error, and Scenario 3 addressed measurement error. Note that Scenarios 1 and 3 might be expected to affect both relative and absolute outcomes, whereas Scenario 2 was designed as a control question because it should have no effect on the outcome.

Phase III Results

For each of the six questions about hypothetical scenarios, the response was "yes" or "no." The answers were classified into one of three categories: (a) yes, with mechanism explanation; (b) yes, without mechanism explanation; and (c) no. Overall coding agreement across all categories was 90%.

When reasoning about the hypothetical scenarios, 100% of the participants correctly said that the ball hitting the side (execution error) and the ball rolling back a few steps (measurement error) could influence how far a ball went and whether the same ball would go farther. Eighty-eight percent were able to offer a mechanistic explanation for why the ball hitting the side of the ramp would make a difference, and 72% were able to offer an explanation for why the ball rolling back a few steps would make a difference. For the control question asking whether the timing of the 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).

A summary variable was created to see if there were differences in ability to recognize different sources of error. For each of the six questions, there were three possible answers about whether the factor would have an effect (*yes with mechanism, yes without mechanism, and no*). For the questions about the ball hitting the side and the ball rolling back, 1 point was added for each yes answer for which a reason was given (up to 4 points, for the two questions about each factor). For the question about the timing of gate release, a point was given for either an answer that it had no effect, or an answer that it had an effect, along with a viable mechanism for the effect. This yielded a 6-point variable.

Overall, 34% of children answered all 6 questions completely and accurately (6 points on the composite score). There were significant grade differences in the distribution of expertise across the levels. Fifty percent of the fourth graders received 6 points, and 22.2% of the second graders received 6 points. The fourth graders had a mean of 5 points, and the second graders had a mean of 4 points, $t(45) = 2.21, p = .03$.

Analysis of variance (ANOVA) tests did not reveal any evidence of a relation between the ability to recognize and explain the role of these error sources (Phase III tasks), ability to design unconfounded experiments (Phase I tasks), or the ability to name other sources of error (in both Phase I and Phase II).

GENERATING ERROR SOURCES ACROSS PHASES

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., when explaining why the balls would not land in the same place, or explaining why a ball rolled down the same ramp five times appeared to have taken a different amount of time for each run). The responses were coded for mention of possible sources of execution error or measurement error, such as the ball hitting the side, or wind blowing, or the stopwatch being started or stopped at the wrong time, as a reason for the variation in results. Ninety percent of children were able to name at least one source of error. The 5 children who did not name any sources of error were all second graders. In addition, the fourth graders named, on average, nearly twice as many error sources [mean for second grade = 3.8, mean for fourth grade = 7.2, $t(46) = 3.31$, $p = .002$].

DISCUSSION

Error can never be eliminated from empirical science: The best we can do is to minimize, recognize, and account for its effects. This task is a constant challenge for practicing scientists, and it is therefore a complex task to teach. Accurate assessment of just what children understand about error is an important first step in planning an effective teaching strategy for beginning science students. The aim of this article is to provide a conceptual structure—a five-stage taxonomy of error—for exploring children's error understanding and to apply it to an example that demonstrates that the taxonomy can be used to develop improved assessments of what children do and do not know about error.

Perhaps our most interesting finding is that although uninstructed elementary school children usually fail to design unconfounded experiments, they do understand quite a bit about error and its ramifications. The discovery of this level of error understanding—even in the limited domain explored here—is both unexpected and important. The finding that second graders are able to propose several different potential influences on experimental outcomes in a familiar domain suggests that they can effectively combine their domain knowledge with their causal reasoning abilities in an experimental context.

Design Error

As found in earlier studies (Chen & Klahr, 1999; Toth et al., 2000), most second and many fourth graders had difficulty designing unconfounded experiments. However, children's ability to allude to aspects of their design in explaining their predictions and justifying their conclusions suggests some understanding of the link between design and outcome, and outcome and conclusions. Although second graders lack full comprehension of the procedural and conceptual knowledge necessary to design unconfounded experiments, they appear to have already acquired a preliminary basis for that understanding. This consistency in children's justifications and conclusions also suggests that they are able to reason causally, at least in this context. They recognize that different designs can cause different outcomes, and that the experimental setup provides a basis for both predicting and explaining experimental outcomes.

Measurement Error

Many children demonstrated evidence of understanding measurement error in the outcome-measurement stage (the simple apparatus used here confined all measurement error to this stage, as there was no measurement done in the setup stage). Most participants could name sources of measurement error as potential reasons for data variation or could recognize measurement error as a possible influence on an experiment's outcome. This type of error was less salient in several of the specific tasks in this study because the distance the balls rolled was measured discretely. In addition, children did not refer to measurement errors when asked to justify their level of confidence in their conclusions. Thus, the extent to which this understanding is integrated with the rest of their knowledge about experimentation, and whether it is used in drawing conclusions from experiments, remains to be investigated.

Execution Error

Children's skill at naming and recognizing many sources of execution-stage error across several of the tasks used in this study indicates that they understand the idea of multiple causality, that is, that many variables can affect an outcome. When explicitly asked about error, children found it easy to propose different possibilities. 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 sureness 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.

Interpretation Error

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 evidence 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 with 95% of their correct predictions. This difference suggests that at least some children are sensitive to conflicts between theory and evidence. Prior domain knowledge appeared to guide their reasoning; children justified most of their predictions by referring to the expected effects of the target and nontarget variables (although this reliance varied based on design).

Children's knowledge of different types of error did not appear to be directly linked to their confidence in drawing conclusions from either confounded or unconfounded experiments. They rarely cited actual or potential errors in design, setup, execution, or measurement as reasons for their confidence or lack of confidence that a factor would have an effect. However, they did name some sources of error as justifications for why explicit empirical results were as they were, and, when probed, they recognized that these factors could play a role.

Several factors might have led children to refer to errors in reasoning in some contexts but not others. One important consideration is that children's relatively good domain knowledge about causal factors in the ramps domain produced a high proportion of accurate predictions. Thus, prior knowledge may have overpowered all other factors and minimized the role of error as causal. Another possibility—as previously discussed briefly—is that although children do understand the different types of error, they have not yet integrated this knowledge with the other information they know about experimentation. Thus, they may not have realized why it might be important to consider potential errors in any earlier stages when drawing conclusions about the effects.

The finding that children differentiate the role of error in relative as compared with absolute measurements suggests a nascent understanding of the difference between overall effects and the specific items that comprise a sample. Children's confidence that the relative ordering would remain the same is an indication that they expect the relative ordering of sample means to remain unchanged, whereas their lack of confidence in absolute outcomes remaining the same indicates their understanding of variability within each sample. The fact that this confidence was linked with the actual difference in the distance the two balls traveled suggests that children are considering effect size in drawing conclusions from comparisons. Our finding that fourth graders distinguish these two situations more clearly than second graders is consistent with other recent work suggesting that 10-year-olds are better than 8-year-olds at differentiating the importance of keeping records of probabilistic versus deterministic events (Rapp & Wilkening, 2001).

Even when they are unable to design unconfounded experiments, second and fourth graders can reason about data variability and alternative causes (error) in interpreting data. However, because the design (whether all relevant variables have been controlled) is considered in the interpretation, it seems necessary that the understanding of one process cannot be completely independent from the understanding of the other. Knowledge about the role of alternative factors might be linked to a basic understanding that it is important to consider many things when deciding what to control. In fact, it seems likely that the explicit training such as that used in Chen and Klahr (1999) was effective precisely because the groundwork was already there: Children knew enough about potential error sources that, once they learned the importance of controlling variables, they were able to learn quickly how to design good experiments.

Connectivity of Types of Error Understanding

Children reasoned about errors in four of the five different stages of experimentation.[7] They considered the validity and informativeness of their designs in choosing their experiments, they considered the role of several errors in execution and measurement as explanations for variation in results—whether observed or hypothetical—and they explained their confidence in their interpretation of the results and their expectations of the outcome of repeated experiments.

However, children's understanding of different types of error appears to follow independent developmental paths—at least over the limited age range studied here. This evidence suggests that children conceptualize the different types of experimental error as distinct and not necessarily related. There were consistent grade differences in performance across tasks but no correlations among measures of different types of error understanding once grade was controlled for. This result is consistent with children's lack of mention of execution and measurement error sources in drawing conclusions about results.

The different error types were expressed in varied ways in this study. Children could make design errors (by setting up a confounded test) or interpretation errors (by drawing incorrect conclusions or offering justifications that sug-

[7] In this study, children did not have to consider measurement error in the setup stage because there were neither independent instruments to calibrate nor independent variables that could be set up incorrectly.

gested a misconception about the logic and process of experimentation). However, for measurement and execution errors, children were asked to generate sources of these errors and to identify them as possible influences on outcomes and conclusions. The differences in the task demands may have led to some differences in the apparent patterns of error understanding. On the other hand, it is also possible that these different forms of error are indeed differentially important, and that in fact, the relative importance of each depends on the domain and the specific task and goal. For science education in particular, it may be important to consider which of these understandings are acquired earliest and without instruction and which ones need to be taught explicitly in elementary school science classes.

Grade Differences

Second and fourth graders performed comparably on many of our assessments, but there were also several tasks on which fourth graders outperformed second graders, including designing unconfounded experiments, naming sources of error, and differentiating between relative and absolute measurements. Although none of the students in this study had received any explicit classroom instruction about designing science experiments and considering variation and error in data, fourth graders have had more experience and general science education. Even on the tasks on which second graders performed well, fourth graders tended to demonstrate more knowledge. For example, whereas most second graders could name at least one source of error throughout the experiment, every fourth grader could do so, and fourth graders, on average, named nearly twice as many error sources.

Causal Reasoning

Children demonstrated some causal reasoning skills and also indicated some difficulties in each stage of the experiment. In one sense, children's understanding of causality is very strong; they readily used prior knowledge to both generate and recognize potential sources of measurement and execution error. In another sense, this understanding of causality is far from complete because children had difficulty when asked to apply this understanding to more complicated situations in which they needed to consider error as one of many possible influences on an outcome. In general, children were much better at generating ideas in the abstract, but they had more difficulty in developing explanations for specific outcomes in considering what had actually occurred during the experiment and how these events might affect conclusions they could draw.

CONCLUSIONS

Taken as a whole, these results tell us that second- and fourth-grade children understand many things about error. They can recognize potential sources of error, they can generate a list of factors that might affect the execution of the experiment and the measurement of results, and they can differentiate the contexts in which error plays a critical role from those in which it is less salient. However, although children do understand a lot about error, they have yet to completely integrate the different fragments of their knowledge about experimental error into a coherent whole. Most of the children who could reason successfully about all of the things listed previously were unable to design unconfounded experiments consistently and rarely referred to potential errors in justifying their conclusions from experiments. Fourth graders were more adept at many of the tasks presented here, suggesting a gradual developmental shift as children learn more about experimentation and causation. These findings suggest that (a) children at these ages still have much to learn about error and its causes, and (b) researchers still have much to learn about what children know about error.

ACKNOWLEDGMENTS

This work was supported in part by grants from NICHHD (HD25211) and the James S. McDonnell Foundation (96–37) to the second author. Thanks to the children, parents, principal, and teachers at St. Teresa of Avila for their participation and cooperation. We also thank Jennifer Schnakenberg for her insightful suggestions at several stages of the project, and for her assistance in coding. We are grateful for constructive comments on earlier drafts from Giora Hon, Brad Morris, Anne Siegel, Vladimir Sloutsky, Eva Toth, Lara Triona, Jolene Watson, and Corinne Zimmerman. Portions of these results were presented at the meeting of the Society for Research in Child Development, April 2001, Minneapolis, MN, the meeting of the American Educational Research Association, April 2001, Seattle, WA, and the meeting of the Cognitive Science Society, August 2001, Edinburgh, Scotland.

REFERENCES

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

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. F. Friedman (Ed.), *The developmental psychology of time* (pp. 209–245). London: Academic Press.

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

Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching, 35,* 623–654.

Doherty, M. E., & Tweney, R. D. (1988). *The role of data and feedback error in inference and prediction* (Final report for ARI Contract MDA903-85-K-0193). Bowling Green, OH: Bowling Green State University.

Dunbar, K. (2001). What scientific thinking reveals about the nature of cognition. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 115–140). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Freedman, E. G. (1992, November). *The effects of possible error and multiple hypotheses on scientific induction.* Paper presented at the meeting of the Psychonomic Society, St. Louis, MO.

Frye, D., Zelazo, P. D., Brooks, P. J., & Samuels, M. C. (1996). Inference and action in early causal reasoning. *Developmental Psychology, 32*, 120–131.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71,* 1205–1222.

Gopnik, A, Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three- and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37,* 620–629.

Gorman, M. E. (1986). How the possibility of error affects falsification on a task that models scientific problem-solving. *British Journal of Psychology, 77,* 85–96.

Gorman, M. E. (1989). Error, falsification and scientific inference: An experimental investigation. *Quarterly Journal of Experimental Psychology, 41A,* 385–412.

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

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* (A. Parsons & S. Milgram, Trans.). New York: Basic Books.

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.

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

Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin, 125,* 524–543.

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

Koslowski, B., & Masnick, A. (2002). Causal reasoning. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 257–281). Oxford, England: Blackwell.

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

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

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

Mellor, D. H. (1967). Imprecision and explanation. *Philosophy of Science, 34,* 1–9.

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development, 5,* 193–207.

O'Connor, R. M., Jr., Doherty, M. E., & Tweney, R. D. (1989). The effects of system failure error on predictions. *Organizational Behavior & Human Decision Processes, 44,* 1–11.

Penner, D., & Klahr, D. (1996). When to trust the evidence: Further investigations of the effects of system error on the Wason 2–4–6 task. *Memory & Cognition, 24,* 655–668.

Petrosino, A. J., Lehrer, R., & Schauble, L. (in press). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning.*

Rapp, A. F., & Wilkening, F. (2001, April). *Children's recognizing of when a protocol is useful: Distinguishing deterministic and probabilistic events.* Poster session presented at the biennial meeting of the Society for Research in Child Development, Minneapolis, MN.

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

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28,* 859–882.

Schlottman, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology, 35,* 303–317.

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

Sheynin, O. B. (1966). Origin of the theory of error. *Nature, 211,* 1003–1004.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*(1, Serial No. 194).

Siegler, R. (1975). Defining the locus of developmental differences in children's causal reasoning. *Journal of Experimental Child Psychology, 20,* 512–525.

Siegler, R. S., & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Developmental Psychology, 10,* 401–402.

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 research-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.