Journal of Cognition and Development

ISSN: 1524-8372 (Print) 1532-7647 (Online) Journal homepage: http://www.tandfonline.com/loi/hjcd20

Data-Driven Belief Revision in Children and Adults

Amy M. Masnick, David Klahr & Erica R. Knowles

To cite this article: Amy M. Masnick, David Klahr & Erica R. Knowles (2017) Data-Driven Belief Revision in Children and Adults, Journal of Cognition and Development, 18:1, 87-109, DOI: 10.1080/15248372.2016.1168824

To link to this article: http://dx.doi.org/10.1080/15248372.2016.1168824

Accepted author version posted online: 12 May 2016.
Published online: 12 May 2016.

JOURNAL OF COGNITION AND DEVELOPMENT, 18(1):87–109

ISSN: 1524-8372 print/1532-7647 online
DOI: 10.1080/15248372.2016.1168824

Routledge
Taylor & Francis Group

Data-Driven Belief Revision in Children and Adults

Amy M. Masnick
Hofstra University

David Klahr
Carnegie Mellon University

Erica R. Knowles
Berklee College of Music

The ability to use numerical evidence to revise beliefs about the physical world is an essential component of scientific reasoning that begins to develop in middle childhood. In 2 studies, we explored how data variability and consistency with participants' initial beliefs about causal factors associated with pendulums affected their ability to revise those beliefs. Children (9–11 years old) and college-aged adults ran experiments in which they generated, recorded, and interpreted data so as to identify factors that might affect the period of a pendulum. In Study 1, several children and most adults used observed evidence to revise their initial understanding, but participants were more likely to change incorrect noncausal beliefs to causal beliefs than the reverse. In Study 2, we oriented participants toward either an "engineering" goal (to get an effect) or a "science" goal (to discover the causal structure of the domain) and presented them with variable data about potentially causal factors. Science goals produced more belief revision than engineering goals. Numerical data, when presented in context, with appropriate structure, can help children and adults reexamine their beliefs and initiate and support the process of conceptual change and robust scientific thinking.

Reasoning about data is a fundamental aspect of scientific thinking (National Research Council [NRC], 1996; Next Generation Science Standards [NGSS], 2013). The NRC's (2007) list of core proficiencies in K–8 science includes students' ability to collect, interpret, and evaluate data. These abilities are also emphasized in the mathematics curriculum proposed by the National Council of Teachers of Mathematics (2000). In this article, we focus on differences in how children and adults reason about physical phenomena as they integrate numerical data from experiments with prior beliefs and how data variability affects this process.

Effective use of data requires understanding how to reason about numerical evidence—particularly when such evidence is not consistent: "Using data from actual investigations from science . . . students encounter all the anomalies of authentic problems—inconsistencies, outliers, and errors—which they might not encounter with contrived textbook data" (NRC, 1996, pp. 217–218). Reasoning about such

Correspondence should be sent to Amy M. Masnick, Psychology Department, 207 Hauser Hall, 135 Hofstra University, Hempstead, NY 11549, USA. E-mail: amy.m.masnick@hofstra.edu
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hjcd.

data involves a) using background theoretical knowledge to situate the numbers (Lovett & Chang, 2007; Schwartz, Sears, & Chang, 2007), b) understanding basic mathematics and knowledge of how empirical data relate to natural phenomena (NRC, 2007), and c) understanding when it is appropriate to revise concepts based on new empirical information (Chinn & Malhotra, 2002; Koslowski, 1996; Kuhn, Amsel, & O'Loughlin, 1988).

Belief Revision

Many aspects of scientific thinking require that students revise and refine not only their understanding of everyday terminology, but also their prior concepts (Vosniadou, 2013). For example, while an untrained child may hold an undifferentiated view of heat and temperature, science instruction aims to clarify and define the distinction between them. In such cases, knowledge is organized based on theories, structured by domain, and knowledge acquisition can be characterized as a change in those structures (Vosniadou & Brewer, 1992).

An important aspect of belief modification (Vosniadou, 2008, 2007) is the extent to which empirical evidence inconsistent with a concept influences its modification. Specifically, what happens when numerical data generated during simple physical experiments are inconsistent with initial concepts? Although some studies have shown that such challenges lead children to modify their beliefs (cf. Burbules & Linn, 1988; Kloos & Somerville, 2001; Limón & Carretero, 1997; Penner & Klahr, 1996), many studies have indicated that empirical contradictions, on their own, are often insufficient to produce such changes, both for children and for adults (e.g., Chinn & Malhotra, 2002; Koslowski, 1996; Kuhn et al., 1988; for detailed overviews, also see Limón, 2002; Zimmerman, 2007). Thus, even though true conceptual change usually occurs only after a protracted period, it is important to identify features of empirical data that initiate the first steps in the conceptual change process.

Prior beliefs, of course, play a crucial role in the interpretation of new data. In reasoning about cause and effect, people use information about covariation (e.g., Cheng, 1997; Tenenbaum, Griffiths, & Kemp, 2006) and causal mechanisms (e.g., Ahn, Kalish, Medin, & Gelman, 1995; Koslowski, 1996; White, 2003), and they combine both covariation and mechanism (e.g., Chinn & Brewer, 2001; Fugelsang & Thompson, 2003). The process of data-driven belief revision about causal factors is asymmetric: It is generally more difficult to revise an erroneous belief that a factor is causal and conclude from new data that it is noncausal than it is to revise an erroneous belief that a factor is noncausal and decide that it *is* causal (Amsel, Goodman, Savoie, & Clark, 1996; Kanari & Millar, 2004).

Although many studies have explored the relationship between belief strength and the likelihood of belief modification, only a few studies have examined characteristics of quantitative evidence that might lead to such change. One way to investigate this relationship is to embed data variation in a larger experimental context. Using this approach, children are asked to collect data and make sense of it, considering real-world measurements with results that are inherently variable. Such studies have shown that children have some understanding that measuring instruments can affect variability of results. For example, Petrosino, Lehrer, and Schauble (2003) found that 10-year-old children expected more variation in measurements when using a hand-crafted cardboard protractor than when using a standardized plastic protractor. In another richly contextualized study, Lehrer and Schauble (2004) examined how 10- to 11-year-old children learned about variation in the growth rate of plants through trying to represent aggregated data and how children's content knowledge informed their

data representation abilities. This approach involves studying variability by situating it in a broader context, with understanding variability as one component of the learning process.

In contrast, more targeted, lab-based studies can complement contextualized studies by focusing on controlled manipulations of variability to look at their effect in isolation. For example, in a lab study, Masnick and Morris (2008) found that even children with no statistical training were more confident about conclusions drawn from data with larger, rather than smaller, sample sizes and from data with less variability. Eight- to 10-year-olds expect more variation in absolute results (the specific distance that a ball rolls down a ramp) than in relative results (which of two ramps will cause the ball to go farther). That is, they understand that although there is variability in the system, it would rarely override the primary causal factors (Masnick & Klahr, 2003). The studies reported in this article approached the issue of statistical variability in laboratory contexts to isolate the effects of manipulated specific factors.

One challenge in making sense of empirical data is interpreting numerical variation. Researchers use inferential statistics to assess the likelihood of patterns occurring by chance. However, such analyses are complex. Although even elementary school students sometimes expect random variation—and reason accordingly—they also sometimes mistake error variance for true effects (Lubben & Millar, 1996; Masnick & Klahr, 2003). In fact, many college students struggle with a full understanding of the variability inherent in scientific measurement (Lubben, Campbell, Buffler, & Allie, 2001), even after taking courses in laboratory science (Konold & Pollatsek, 2002; Volkwyn, Allie, Buffler, & Lubben, 2008) or statistics (delMas & Liu, 2007).

The Pendulum Context

The pendulum context is a classic one in the study of cognitive development (e.g., Inhelder & Piaget, 1955/1958), and it is also a widely used vehicle for teaching about experimentation in the elementary science classroom (NRC, 1996). Pendulum experiments provide a good physical context for studying the interaction between prior beliefs (both correct and incorrect) and empirical numerical evidence. The period of oscillation of a simple pendulum is affected by its length (shorter pendulums swing faster), but contrary to a widespread misconception, it is not affected by the mass of the bob. Other factors such as air currents, extreme height at release, a push at release, or variation in the gravitational constant (distance from the earth) can also affect a pendulum's period of oscillation. When children and adults are asked to suggest which variables determine the period of a pendulum, they usually propose length, correctly, and several additional variables—such as the mass of the bob or the amplitude of oscillation—that in fact do *not* matter in most constrained laboratory contexts in which the pendulums are tested and explored (any effects that the noncausal variables appear to produce are error variance; Frick, Huber, Reips, & Krist, 2005). However, in real-world experimentation, repeated measurements usually yield slightly different results. This error variance, as well as the error introduced by various measurement schemes commonly used in the science classroom (typically, timing a fixed number of swings with a handheld stopwatch and then computing the mean time per swing), may give the appearance of some systematic causal factor. Here, repeated measurements are likely to vary slightly when comparing factors other than string length, and they vary more dramatically when comparing swings of the pendulum with different string lengths. Thus, the pendulum provides a potentially rich and informative context in which to study children's understanding and use of varied numerical data.

In this article, we investigated the extent to which variance in data influences the interpretation of those data when they conflict with conceptual understanding. More specifically, we assessed children's and adults' ability to conduct and interpret experiments in a domain in which most of the putative causes do *not* affect the outcome, but in which error variance may obscure that fact. Here, we are interested in whether varied data are one type of evidence that can initiate changes in beliefs and whether the systematicity of the testing approach influences how numerical evidence is used. We hypothesized that seeing the wide differences in numerical values representing a true difference, as compared to the smaller differences in numerical values representing error variance, would make it more likely that children and adults would recognize which factors were causal. We also hypothesized that the systematic, scientific comparison of different variables would lead to improved knowledge about causal and noncausal factors and that this scientific approach would be more effective in inducing belief revision than a more goal-oriented, engineering approach, aiming to achieve a specific effect. The National Science Education Standards (NRC, 1996) are broken into recommendations for kindergarten through fourth grade and for fifth through eighth grade, so the fourth- and fifth-grade students participating in the present research are on the cusp of a transition in which they develop a more sophisticated understanding of science and the nature of evidence, and thus, they are a particularly informative age group to study. We hypothesized that some fourth and fifth graders could learn with guidance but that they would have more difficulty than adults, given their less broad prior knowledge and weaker ability to fit in new evidence with existing beliefs. At the same time, many adults still struggle with these misconceptions, and thus, we chose to compare adults who share some misconceptions with children to see whether the effects of systematic testing differ across age groups.

STUDY 1

Method

Participants. Participants included 49 children and 28 adults. Children were fourth and fifth graders (M_{age} = 10;9, SD = 7.32 months; 43% male) from a private college-preparatory urban school and were recruited from letters sent to parents. Adults were undergraduates of typical college age (58% male) participating in exchange for course credit.[1]

Design. The specific pendulum attribute that the participant investigated (length, weight, and angle) was a within-participants factor, with every participant comparing several trials at each level of these three variables. Phase (pretest, test, and posttest) was also a within-participants factor. Dependent measures were accuracy and confidence.

Materials. Pendulums with three binary dimensions (bob mass: heavy/light; string length: long/short; and angle of release: high/low) were assembled during the experiment

[1] It is possible that there might be substantial differences in the two populations other than age (e.g., scientific aptitude and interests, etc.). However, based on our knowledge about the educational pathways of the children and adults in these two studies, we are confident that overall intelligence and motivational levels of the two populations are unlikely to be confounding factors.

and suspended from a wooden frame with a hook (see Figure 1). Each pendulum was constructed by choosing one of two different bobs and one of two different string lengths. The bob was then released from one of two different angles from the perpendicular (the two angles were both low enough that in practice, they had no effect on pendulum speed). At the outset of the study, two different bobs and strings were placed in front of the rack, as was a "spacer," to be used to set the high- or low-release angles. The bobs were fashioned from 170.1 g glass jars filled with different amounts of colored sand. The jars were sealed shut and had hooks attached to the cap. The light jar (168 g) had some colored sand in it, and the heavy jar (326 g) was full of colored sand to make a clear visual distinction of mass. The heavy jar also had a small weight hidden inside, leading to a highly salient difference in weight as participants handled the jars in setting up the pendulums. When setting up a pendulum, participants picked a string length, hung the string on the pendulum frame hook, and then picked a jar (heavy or light) to hook to the bottom of the string. Then, they set one of the two release angles by aligning the free end of the pendulum with the spacer, oriented either with its long side horizontal and its short side vertical to yield a low angle from the top of the spacer (low release) or its short side horizontal (and long side vertical) to yield a high angle (high release).

Procedure. All participants were interviewed individually, and their beliefs about what factors affect pendulum swing timing were assessed at three points: a) at pretest, before they ran

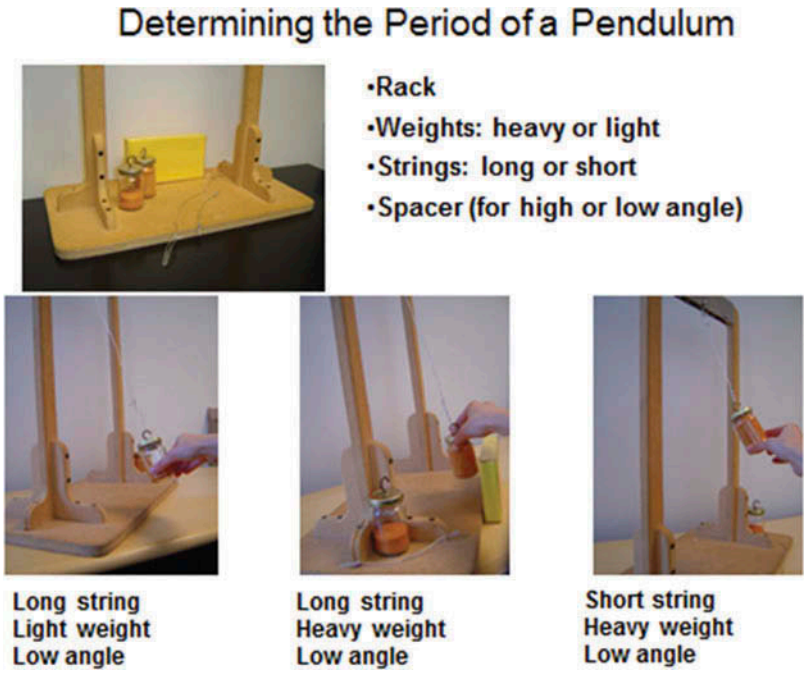

FIGURE 1. Pendulum setup for Study 1, with three dichotomous variables of bob weight, string length, and release angle.

any experiments; b) at test, immediately following the data generation process during which they had "generated" and recorded the set of data points for each of the possible factors; and c) at posttest, following all experimental runs, when they were asked to review the results of their experiments and state their final beliefs.

In the pretest phase, participants were asked whether they thought each variable (string length, bob mass, and release angle) made a difference in how fast a pendulum went and, if so, about the direction of the effect (e.g., they might assert that pendulums with short strings go faster than pendulums with long strings). Then they ranked their confidence on a 4-point scale: "totally sure," "pretty sure," kind of sure," and "not so sure."

During the test phase, participants timed the period of pendulums having different string lengths, bob masses, and starting angles. For each configuration of these three potentially causal variables, one was varied while the other two were held constant.[2] Because we wanted to examine participants' ability to calibrate high versus low variability in this context, for half of the participants, length was the first factor to be explored (producing relatively large differences between outcome values for the initial comparisons), and for the other half of the participants, length was the final factor to be explored (producing relatively low variability for the initial comparisons that varied bob mass and starting angle). Participants were asked to time 10 swings of a pendulum in each of two configurations (e.g., 10 swings with a heavy bob and 10 swings with a light bob), while string length and release angle were held constant. The participant released the pendulum and counted 10 swings, and the experimenter read out the "resulting" time. For each variable under investigation, this procedure was repeated eight times: four for one level of the variable (e.g., a long string) and four for the other level (e.g., a short string). Unknown to the participant, all times read by the experimenter were predetermined and not actual measures of the current trials. This manipulation ensured that each participant was exposed to exactly the same data variation. The times that were falsely presented as veridical were entirely plausible because they were very close to the actual times for each pendulum trial. (Debriefings revealed that none of the participants suspected our manipulation.) Participants recorded the times read out on a worksheet that was provided, so that during their session they could always view the full set of times from all the experiments they had run. After eight test trials for each variable, they were asked their beliefs about that variable, whether it made a difference (if so, in which direction), and their confidence. They then went on to test the next variable.

All of the data sets varied slightly: For each given variable, there were no two measurements announced to participants that were identical within the set of trials. Thus, sometimes the values were in the predicted direction and sometimes they were not, but there were only small differences between the values. However, because string length was the only variable that caused a true effect, the difference in times read from the trials with a short string and those read from the trials with a long string was much more pronounced. The data from these two sets of trials did not overlap (mean duration of 10 swings of the pendulum for long-string trials = 14.2 s, SD = 0.22s; mean duration for short-string trials = 11.3 s, SD = 0.39 s). For the heavy/light-weight trials and the high/low-release trials, the mean times were identical across

[2] Thus, all of the experiments that our participants observed were unconfounded. Their ability to create unconfounded experiments (Chen & Klahr, 1999; Klahr & Nigam, 2004) was not assessed in this study.

the two conditions being compared (14.15 s in the weight trials and 14.45 s in the angle trials). As noted, each participant experimented with the effects of length, weight, and angle.

In the posttest, participants were asked a final set of questions to provide their beliefs about whether each variable made a difference, and they rated their confidence in their beliefs.

Results

First, we examined whether participants correctly assessed the causal effect of each variable (length, angle, and weight) at pretest, test, and posttest and whether this effect differed by age. Causal assessment accuracy was operationalized by participants' answers to the question, "Does X make a difference?" for each variable. Correct answers to the length question had to assert that length made a difference *and* that the pendulum with the short string went faster, and answers to the angle and weight questions had to assert that these variables made no difference in pendulum speed. Adults and children started out with very similar scores (see Figure 2 for totals).[3] At pretest, most participants believed (correctly) that length was causal (85.4% of the children and 73.1% of the adults), and most participants believed (incorrectly) that weight and angle also were causal (87.5% of children and 80.8% of adults believed that weight was causal, and 95.8% of children and 61.5% of adults believed that angle was causal).

Both adults and children learned from running the experiments, and there was no effect of whether the length variable was presented first or last. Both adults and children showed significant gains between pretest and test phases, and both maintained their test phase scores through the posttest. However, adults learned significantly more about the causal and noncausal factors. Mauchly's test indicated that the sphericity assumption was violated, $\chi^2(2) = 17.688$, $p < .001$, so the Huynh-Feldt correction was applied to adjust degrees of freedom for analyses with the repeated variable of phase ($\varepsilon = .867$). A 3 (phase) × 2 (age group) × 2 (order) analysis of variance (ANOVA) with total number of correct answers (range = 0–3) as the dependent variable showed a main effect of phase, with means (*SD*s) of 1.11 (0.61), 2.07 (0.89), and 2.14 (0.90) across pretest, test, and posttest, respectively, $F(1.73, 119.64) = 98.69$, $p < .001$, partial $\eta^2 = .59$; a main effect of age with average means of 1.72 ($SD = 0.83$) for children and 2.89 ($SD = 0.43$) for adults, $F(1, 69) = 53.28$, $p < .001$, partial $\eta^2 = .44$; and no main effect of order ($p = .33$). There was a significant Phase × Age interaction, $F(1.73, 119.64) = 15.02$, $p < .001$, partial $\eta^2 = .18$, a nonsignificant Phase × Order interaction ($p = .553$), and a nonsignificant Age × Order interaction ($p = .73$). Simple main-effects analyses for the Phase × Age interaction with a Bonferroni correction demonstrated that at pretest, adults were a little more knowledgeable than the children: mean correct for adults = 1.31 ($SD = 0.74$ versus mean correct for children = 1.0 ($SD = 0.51$), $F(1, 69) = 4.53$, $p = .037$, partial $\eta^2 = .06$. However, at test, they were further apart, with a much larger effect size: 1.64 ($SD = 0.76$) versus 2.85 ($SD = 0.46$), respectively, $F(1, 69) = 58.83$, $p < .001$, partial $\eta^2 = .44$. There was a similar pattern at posttest: 1.72 ($SD = 0.83$) versus 2.88 ($SD = 0.43$), $F(1, 69) = 43.65$, $p < .001$, partial $\eta^2 = .39$. There was also a significant three-way interaction between phase, age, and order, $F(1.73, 119.64) = 3.38$, $p = .044$, partial $\eta^2 = .05$. However, simple-effects contrasts with a Bonferroni correction, comparing length first versus length last for each age group and phase, revealed only one marginally significant effect, a difference for adults at pretest, whereby those adults who ended up seeing length as the first of the three variables tested had less initial knowledge ($M = 1.08$, $SD = 0.64$) than those who saw

[3] One child was correct on all three variables at pretest and was eliminated from further analyses.

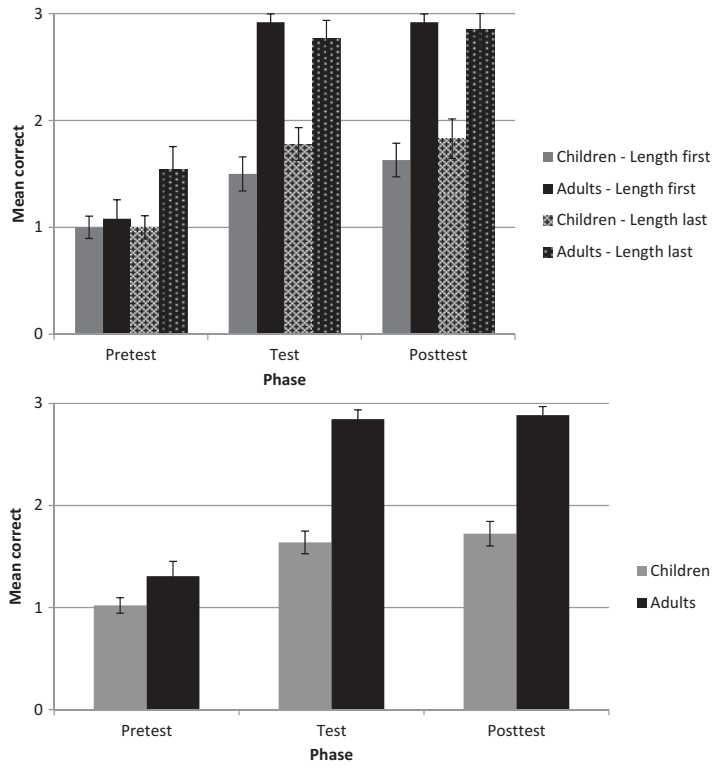

FIGURE 2. Study 1: A) The average number of variables (out of three) for which children and adults correctly assessed the causal role at pretest, test, and posttest by order of first variable. B) The average number of variables (out of three) for which children and adults correctly assessed the causal role at pretest, test, and posttest, collapsed across order. (Error bars represent standard errors.) There were significant effects of phase, of age, and of the Phase × Age interaction (all ps < .001).

length as the last of the three variables tested (M = 1.54, SD = 0.78), $F(1, 69) = 3.96$, $p = .051$, partial $\eta^2 = .05$. No other contrasts approached significance (see footnote 4 for details).[4]

Most children and adults learned that length made a difference. Although children's gains from pretest to posttest were statistically significant, their accuracy on weight and angle remained at very low levels (see Table 1 for a full set of accuracy levels). We also determined which participants became pendulum "experts" who correctly identified length as the only causal factor. At test, only 14.6% of children were experts compared with 88.5% of the adults, $\chi^2(1, N = 74) = 38.19$, $p < .001$; and by posttest, 22.9% of children were experts compared with 92.3% of adults, a significant age difference, $\chi^2(1, N = 74) = 32.58$, $p < .001$.

[4] For children, there were no differences between the group that saw length as the first variable at pretest ($p = 1.0$), at test ($p = .153$), or at posttest ($p = .340$). For adults, there were no differences between the groups at test ($p = .561$) or posttest ($p = .785$).

TABLE 1.
Accuracy levels for each variable in Study 1, by age group

	Length			*Weight*			*Angle*		
	Pretest	*Test*	*Posttest*	*Pretest*	*Test*	*Posttest*	*Pretest*	*Test*	*Posttest*
Children	85.7%	98%	100%	13.4%	32.7%	34.5%	6.1%	37.5%	41.7%
Adults	73.1%	100%	100%	19.2%	88.5%	92.3%	38.5%	96.2%	96.2%

Participants also assessed their confidence on a 4-point scale, from 0 as not sure to 3 as totally sure. We used this scale as a dependent measure and examined how confidence was related to accuracy. For all three variables (length, weight, and angle), at pretest, there were no significant effects of either accuracy or age; everyone was about equally confident, on average. We could not look at length at test and posttest, because nearly every participant was accurate by test in assessing the role of length. However, in assessing the role of weight, there was a main effect of accuracy, $F(1, 69) = 6.37$, $p = .014$, partial $\eta^2 = .08$, though no main effect of age ($p = .489$). There was also an interaction between age and accuracy, $F(1, 69) = 12.67$, $p = .001$, partial $\eta^2 = .16$. See Table 2 for details. Simple-effects contrasts with a Bonferroni correction indicated that children were about equally confident in their assessment of weight, regardless of whether they judged it correctly or incorrectly. In contrast, adults were much more confident when they were correct than when they were not, though only 3 adults gave an incorrect answer. A similar pattern was found at posttest (see Table 2). Similarly, in the test phase for the angle variable, there was a main effect of accuracy, $F(1, 67) = 4.63$, $p = .035$, partial $\eta^2 = .07$, and no main effect of age ($p = .493$). Again, there was an Age × Accuracy interaction, $F(1, 67) = 4.33$, $p = .04$, partial $\eta^2 = .06$. See Table 2 for details. Simple-effects contrasts indicated that children were equally confident whether they correctly assessed the role of angle or incorrectly assessed it. In contrast, the 1 adult who was incorrect about angle at test was not very confident (1.0, kind of sure), while the 25 adults who were correct were far more confident ($M = 2.48$, $SD = 0.65$). A similar pattern was found at posttest.

TABLE 2.
Confidence for each noncausal variable, by accuracy and age

Weight		N *correct*	*Mean confidence* (SD)	N *incorrect*	*Mean confidence* (SD)	*Contrast comparing confidence across accuracy levels*
Test	Children	15	1.53 (0.64)	32	1.81 (0.69)	$F(1, 69) = 1.30$, $p = .26$, partial $\eta^2 = .02$.
	Adults	23	2.30 (0.97)	3	0.67 (0.58)	$F(1, 69) = 11.65$, $p = .001$, partial $\eta^2 = .14$
Posttest	Children	16	1.88 (0.66)	32	2.00 (0.63)	$F(1, 70) = 0.42$, $p = .520$, partial $\eta^2 = .01$
	Adults	24	2.63 (0.58)	2	1.50 (0.71)	$F(1, 70) = 5.92$, $p = .018$, partial $\eta^2 = .08$
Angle						
Test	Children	16	2.00 (0.73)	29	1.97 (0.63)	$F(1, 67) = 0.02$, $p = .87$, partial $\eta^2 = 0$
	Adults	25	2.48 (0.65)	1	1.00	—
Posttest	Children	19	2.05 (0.78)	28	2.21 (0.74)	$F(1, 69) = 0.54$, $p = .47$, partial $\eta^2 = .01$
	Adults	25	2.48 (0.65)	1	1.00	—

Discussion

The results of this study revealed similarities in children's and adults' initial misconceptions, contrasted with differences in their ability to revise those misconceptions in the face of empirical results from experiments. However, when faced with variable data from unconfounded experiments that did not conform to prior expectations, adults were much more likely than children to use those data to revise their beliefs. Nearly all adults—but only a few children—clearly differentiated the small variation in measured outcomes associated with noncausal factors (different weights and angles) from the large variation evident in the measurement of the one variable that did make a difference (length of the string).

Not only were children less likely than adults to update their beliefs, but their confidence was also uncorrelated with their accuracy. Interpreting quantitative data involves understanding not only the empirical evidence observed, but also the implications of that evidence for one's prior beliefs. Evidence that does not match beliefs needs to be explained, usually either by updating the beliefs or by discounting the evidence in some way (Chinn & Brewer, 1998; Chinn & Malhotra, 2002; Penner & Klahr, 1996). For example, the outcome variability in the noncausal variables (angle and weight), even though much smaller than the differences caused by the two levels of the causal variable (length), may have enabled children to reconcile the new evidence with their faulty beliefs. However, seeing the large speed differences between the short and long string before seeing the smaller speed differences between levels of the noncausal variables did not influence conclusions, suggesting that the differences between levels of the first variable did not serve an anchoring role. The lack of anchoring may be because of the small amounts of variation or because each variable was assessed in isolation rather than in contrast to the other variables. As Vosniadou (2013) has suggested, the updating of beliefs is only the first step in real conceptual change, whereas the deeper understanding engendered by true conceptual change may take longer and may require repeated exposure.

STUDY 2

Prior research on children's scientific reasoning processes has uncovered two distinct approaches to the task of generating and interpreting evidence; these approaches can affect the extent to which children undergo conceptual change. In a *scientific approach*, variables are tested systematically to identify causal factors, and in an *engineering approach*, children combine the values of potentially causal factors to achieve a specific outcome, such as building the fastest race car (Schauble, Klopfer, & Raghavan, 1991) or constructing a ramp that will make a ball roll the farthest (Siler & Klahr, 2012). Tasks that require more exploration can impose a larger cognitive load than more structured tasks and require more information to be kept in working memory (Kirschner, Sweller, & Clark, 2006). Thus, more exploratory engineering tasks may be more challenging than the guided structure of the systematic comparisons in a science context. If these exploratory tasks are too overwhelming, they may be ineffective in leading to conceptual change.

Engineering goals also lead to actions based on one's prior knowledge (e.g., the design of a fast pendulum will depend on the child's beliefs about the values of the factors believed to be causal; Siler, Klahr, & Matlen, 2013). Relatedly, there is evidence that preschool children tend to explore tasks more thoroughly when they are trying to learn about causal factors (Schulz & Bonawitz, 2007) and that

children explore tasks differentially based on whether the evidence they have seen conflicts with prior beliefs or not (Bonawitz, van Schijndel, Friel, & Schulz, 2012). The behavior of 6- to 7-year-old children suggests that they spend more time exploring when the evidence violates their prior beliefs than when it does not. From this perspective, one might expect more trials of different types when evidence conflicts with prior beliefs.

In contrast, in a "pure" scientific task, setting up a good comparison does not depend on prior knowledge about the system (it does not preclude having such knowledge, but the knowledge should not affect the systematic testing of variables). When given the choice of which type of goal to choose, most third graders choose engineering goals, while fifth graders with high socioeconomic status (SES) like those in our study (but not low-SES students) tend to choose science goals (Siler et al., 2013). If, instead of allowing children to choose their goals, we orient them toward different goals (i.e., engineering or science), how might that orientation influence children's and adults' ability to utilize numerical evidence in reasoning? In turn, how might such a guided goal orientation affect the likelihood of belief modification? In Study 2, we examined how children and adults reason when given an engineering goal as compared to a science goal.

We also wanted to learn more about why children and adults sometimes resist changing beliefs after seeing new data that contradict their prior beliefs. Some research has suggested that in these circumstances, the most likely source of the errors is incorrectly observing evidence when it does not match prior expectations (Allen, 2010; Chinn & Malhotra, 2002). Further, merely seeing an effect may not be enough information to allow children to tie this knowledge with their prior beliefs, and further guidance may aid in this process (Kloos & Somerville, 2001). Thus, the current study included a comparison of participants in two conditions, half of whom were asked to focus on the data generated to highlight what was seen in the data and make the observation as clear-cut as possible and half whose attention was not drawn to the data in this way. Our goal was to build on prior conceptual change research, which suggests that evidence of flawed beliefs can prompt reconsideration of the beliefs.

Method

Participants. Participants included 59 children and 75 adults. The children were fourth and fifth graders from an affluent public suburban elementary school (M_{age} = 10;0, SD = 8.28 months; 54% female) who were recruited from letters sent to parents. The adults were undergraduate students (M_{age} = 19;0, SD = 17.88 months; 59% female) participating in exchange for course credit. One adult who began the study did not complete it.

Design. The design was a 2 (engineering/science) × 2 (attention to data/no attention) × 2 (child/adult) between-participants design. Half the participants were randomly assigned to the science condition, and half were assigned to the engineering condition; within each condition, half of the participants were assigned to the "attention to data" condition and half were assigned to the "no attention to data" condition. There were 19 adults and 15 children in the science attention condition, 18 adults and 15 children in the science no-attention condition, 18 adults and 17 children in the engineering attention condition, and 20 adults and 12 children in the engineering no-attention condition. Within the science condition, there were three variables to be tested, and participants were randomly assigned to one of the three orders of variable testing, determined using a Latin Square design. Thus, there were three orders of testing string length,

bob weight, and string material (see the Materials and Procedure section under Study 2). The dependent measures were accuracy and confidence. We also examined the types of qualitative explanations participants gave for their responses in both conditions.

Materials and Procedure. Sessions were videotaped if the adult participants agreed or if parents agreed to have their child's session videotaped. The sessions of two college students and eight children were not videotaped by request. For these participants, analyses were based on notes recorded during data collection (note that for all participants there were two experimenters present—one person interacted with the participant and the other recorded responses).

There were three phases to the study. Phase I was a pretest to assess all participants' prior knowledge and was identical for the science and engineering conditions. Participants were introduced to the pendulum setup, which involved the same frame and weighted jars as in Study 1 and which used three potentially causal factors, each having two values, as in Study 1. However, for Study 2, we replaced one of the noncausal factors used in Study 1 (angle) with a different noncausal factor: type of the "string." We used four "strings"—two made of wire and two made of chain, with a short (12.7 cm) and long (31.75 cm) version of each (see Figure 3). The purpose of this change in noncausal variable was to slightly extend the generality of our findings by adding a variable we assumed most participants would be less likely to perceive as causal. We were interested in seeing whether the small variability in timing for a variable that had a less obvious mechanism for effect would change beliefs in the same way as for a variable most believed to be causal (weight). The materials were arranged in front of the pendulum (long and short wire and chain strings and heavy and light jars). To familiarize participants with the general procedures, a demonstration pendulum was set up with a yellow string (of different length and material than the test strings) and a wooden ball as a bob. Participants ran a practice trial to allow the demonstration pendulum to swing for five full swings while the experimenter timed the event with a stopwatch.

Participants were then asked pretest questions about factors that could affect a pendulum's speed. For length, they were asked if they thought string length would make a difference in how fast a pendulum swings and to rate how sure they were of this assessment (totally sure, pretty sure, kind of sure, not so sure). Next, they were asked the more specific question of whether they expected the long string to make the pendulum swing faster, the short string to make it swing faster, or neither to affect its speed. If they said they thought one string would make it swing faster, they were then asked whether they thought it would make the pendulum swing a little bit faster, a medium amount faster, or a whole lot faster. An analogous set of questions was asked about the other two variables, bob weight and string material. The demonstration pendulum was removed from the hook at the end of the pretest.

In the test phase, participants in the science condition were asked to test the first variable (length, weight, or material). When testing length, they were asked to set up the pendulum first with a short string made of wire and a heavy bob. They then timed the pendulum for 10 swings and recorded the results on a worksheet provided by the experimenter. Next, participants timed 10 swings of the pendulum with a long string made of wire and a heavy bob. They continued to switch back and forth, changing only string length, for a total of four trials of 10 swings at each string length. After the eight trials, participants were asked whether they thought the long string or the short string made the pendulum swing faster or if they thought both made it swing about the same speed. They were also asked how sure they were of their answer and to provide an open-ended justification for their answer. In this test phase, they repeated this process for each of the three variables.

As in Study 1, the experimenter used a stopwatch that started and stopped with the pendulum swinging, but every participant was read the same times to record on the worksheet and thus saw times that were plausible but not derived from the current trial. No participant indicated suspicion of our deception. Mean times were 8.7 s for the short string and 12.6 s for the long string with both standard deviations of 0.18. For the weight and materials comparisons, the means and standard deviations of the sets of the times were identical over the four trials, all tested with short strings (M = 8.55 s, SD = 0.26), though no values were repeated exactly among the eight "times" participants saw.

In the attention-to-data group in the science condition, participants were asked which pendulum combination went faster and to estimate how much faster the combination went after the first two trials. After each subsequent pair of trials, participants were asked to look at the worksheet and say how many times the short string (or heavy weight/wire string) had gone faster and then how many times the long string (or light weight/chain string) had gone faster. They were then asked to say which combination made the pendulum swing faster overall, to rate how sure they were, and to say how much faster the one variable level went, if they said one went faster. Beyond these additional questions, the attention-to-data condition was identical to the no-attention-to-data condition in the science condition.

In the testing phase of the engineering condition, participants were asked to build a pendulum that went as fast as possible. Participants were told they were going to test the pendulum eight times and could set it up each time any way they thought would make the pendulum go as fast as possible. After setting up each combination, participants recorded the combination used on a worksheet provided by the experimenter, and then the pendulum was timed for 10 trials. As in the science condition, the times the experimenter stated were predetermined. Because string length was the only variable manipulated that affects the speed, the experimenter had a list of times for combinations with long strings and a list for combinations with short strings, and the times participants were given by the experimenter were based on whether they used a long or short string, regardless of the settings for bob weight and string material used. Times for short-string trials were drawn from a data set with a mean of 8.55 s, and times for long-string trials were drawn from a data set with a mean of 12.55 s, both with. a standard deviation of 0.24 s.

After all eight trials, participants were asked which combination made the pendulum go fastest and to explain why they thought it went fastest. Next, participants were asked to build a pendulum that would go as slowly as possible. Again, they tested and "timed" eight trials, with the experimenter reading predetermined times. After eight trials, participants were asked to identify which combination made the pendulum go slowest and why.

In the attention-to-data group in the engineering condition, participants set up the fastest and slowest pendulums in the same way as in the no-attention group. However, after completing the first two timed trials, they were asked to state which one went faster and what was different about the two combinations. After each subsequent trial, participants were asked if the latest trial went faster than the previous trial and, if so, how much faster. After the fourth, sixth, and eighth trials, participants were asked to look at all the times on the sheet and note which ones were the fastest so far and what the combinations were that went fast. After the eighth trial, participants were also asked to identify the combinations that had gone the fastest and then to state what they thought made the fast ones go faster than the slow ones. The analogous procedure was followed for the trials to create the slowest possible pendulum.

In the posttest phase, all participants answered the same series of questions. Each participant was asked again about each of the three variables (in the same order in which they were asked about them in the pretest). For each variable, they were asked if they thought the variable affected the speed of the pendulum and to rate how sure they were of this assessment. They were also asked about the directionality of that effect (i.e., whether they thought one level of the variable [e.g., long string] makes the pendulum go faster, slower, or the same speed as the other level of the variable [e.g., the short string]). Again, they were asked to rate how sure they were. Finally, if participants said a variable affected pendulum speed, they were asked about the relative magnitude of that effect (e.g., did the short string make the pendulum swing a little bit faster, a medium amount faster, or a whole lot faster?).

Results

Recall that in both the pretest and posttest, participants were asked several questions about each variable. For example, they were asked whether length makes a difference, and then regardless of the answer to this question, they were asked if the long string would make the pendulum go faster, the short string would make it go faster, or both would make it go about the same. We measured consistency on these two assessments of beliefs. A consistent reply on the length questions included stating that length makes a difference *and* that the short string makes it go faster or stating that length does not make a difference *and* that they both go about the same speed. An inconsistent reply might be to state that length does not make a difference but that the short string makes the pendulum swing faster. Although 84% of adults were completely consistent in answering questions about all three variables at pretest, only 50% of children were. There was some improvement at posttest, where 87% of adults and 67% of children were completely consistent. However, 33% of children who were consistent at pretest were not consistent at posttest (11% of adults followed this pattern). These results suggest that a) many children are struggling not only in understanding the content knowledge but also in expressing their beliefs clearly, or b) the demand characteristics of an identically repeated question by the experimenter suggest that their initial answer needed to be changed. This issue is addressed in more detail in the Discussion.

Accuracy. One measure of expertise was a measure of accuracy across all three variables. We characterized as "grand experts" all participants who stated accurately that the short string made the pendulum swing faster and that the bob weight and string material did not affect its speed. To be considered a grand expert required both correct answers and consistency across forms of the question (e.g., answering both that material did not make a difference and that the wire string and the chain string would make the pendulum swing at about the same speed). Thus, participants needed to answer six questions correctly to achieve this designation, a higher bar than our definition of expertise required in Study 1. At pretest, no children and only two of the adults were grand experts, and those two adults were eliminated from further analyses. At posttest, 20.3% of the children and 46.6% of the adults were grand experts, a significant age difference, $\chi^2(1, N = 132) = 9.89$, $p = .002$.[5]

[5] Although the criteria for grand expertise were higher here than in Experiment 1, in practice, the results were identical when grand expertise was assessed for Experiment 2 using the criteria of Experiment 1.

We next looked at which factors in our procedure affected participants' accuracy in assessing which variables are causal. A 2 (phase: pretest vs. posttest) × 2 (age group) × 2 (condition: science/engineering) × 2 (attention to data/no attention) mixed ANOVA was run to examine the effect of each variable on accuracy, defined here as the number of variables on which a participant was an expert (range = 0–3). There was a main effect of phase, such that participants were significantly more accurate at posttest than at pretest with 0.75 accuracy ($SD = 0.71$) at pretest as compared with 1.95 ($SD = 0.88$) at posttest, $F(1, 124) = 181.10$, $p < .001$; partial $\eta^2 = .59$; a main effect of age, with adults ($M = 1.51$, $SD = 0.51$) more accurate than children ($M = 1.16$, $SD = 0.52$), $F(1, 124) = 14.38$, $p < .001$, partial $\eta^2 = .10$; and a main effect of condition, with those in the science condition being more accurate ($M = 1.52$, $SD = 0.52$) than those in the engineering condition ($M = 1.15$, $SD = 0.52$), $F(1, 124) = 17.08$, $p < .001$, partial $\eta^2 = .12$. Further, the Phase × Condition interaction was significant, $F(1, 124) = 32.99$, $p < .001$, partial $\eta^2 = .21$; simple-effects contrasts with a Bonferroni correction indicated no difference between conditions at pretest (science accuracy = 0.68, $SD = 0.70$; engineering accuracy = 0.81, $SD = 0.73$, $p = .321$), but there was a significantly higher accuracy rate at posttest in the science condition ($M = 2.36$, $SD = 0.80$) as compared with the engineering condition ($M = 1.48$, $SD = 0.71$), $F(1, 124) = 48.48$, $p < .001$, partial $\eta^2 = .28$. In addition, the Phase × Age interaction was marginally significant, $F(1, 124) = 3.09$, $p = .081$, partial $\eta^2 = .02$. Simple-effects contrasts with a Bonferroni correction indicated no significant difference between the accuracy levels at pretest between children ($M = 0.65$, $SD = 0.71$) and adults ($M = 0.84$, $SD = 0.71$), $F(1, 124) = 2.27$, $p = .135$, partial $\eta^2 = .02$, but greater accuracy for the adults at posttest ($M = 2.17$, $SD = 0.71$) than for the children ($M = 1.67$, $SD = 0.82$), $F(1, 124) = 15.59$, $p < .001$, partial $\eta^2 = .11$. There were no main effects of the attention/no-attention manipulation ($p = .93$) and no significant interactions with any of the other variables. Therefore, data from the attention/no-attention conditions were collapsed in Figure 3, where means are shown.

Goal type (engineering or science) was the most critical factor in whether participants demonstrated belief revision in understanding the role of the noncausal variables: weight and material. At posttest, far more participants in the science condition (in both age groups) than in the engineering condition said that these variables did not matter: Of those who were "weight experts" by posttest (41.7% of participants), 83.6% were in the science condition and 16.4% were in the engineering condition, $\chi^2(1, N = 132) = 42.7$, $p < .001$. Of those who were "material experts" by posttest (56.8% of participants), 65.3% were in the Science condition, $\chi^2(1, N = 132) = 16.3$, $p < .001$. It was also much easier for people to change beliefs from noncausal or inconsistent to consistently causal than from causal or inconsistent to consistently noncausal. Of the 80 participants who were not length experts at pretest, 76 (95%) were experts by the posttest. In contrast, of the 126 participants who were not weight experts at pretest, only 52 (41.3%) became experts by posttest, and of the 91 participants who were not material experts at pretest, 38 (52.7%) became experts by posttest (in both cases, this change involved moving from causal or inconsistently causal beliefs to consistently noncausal beliefs).

Participants also rated their confidence in their beliefs at the pretest and posttest phases. With length, only four participants were incorrect at posttest, so it was not reasonable to compare the confidence of those who were accurate to that of those who were not. When examining age group, condition (science or engineering), and accuracy about the effect of weight, there was an interaction between condition (science or engineering) and accuracy on confidence, $F(1, 124) = 8.37$, $p = .005$, partial $\eta^2 = .06$. Simple-effects analyses with a Bonferroni correction indicated that participants in the science condition were significantly more confident when correctly stating that weight did not

FIGURE 3. Pendulum materials for Study 2, with four strings of two lengths and two materials plus two jars of different weights.

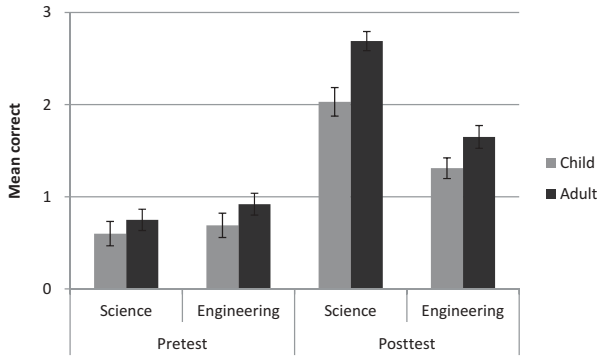

FIGURE 4. Study 2. Mean number of variables (out of three) for which the causal role was correctly assessed, at pretest and posttest, by condition and age group. (Error bars represent standard errors.) There were main effects of condition (science vs. engineering), phase, and age (all *p*s < .001), a Phase × Condition interaction ($p < .001$), and a marginally significant Phase × Age interaction ($p = .081$).

make a difference ($N = 46$, $M = 2.35$, $SD = 0.57$) than when incorrectly stating that it did ($N = 20$, $M = 1.94$, $SD = 0.61$), $F(1, 124) = 5.38$, $p = .022$, partial $\eta^2 = .04$. There was a marginally significant effect of age group, such that children were more confident on average ($M = 2.34$, $SD = 0.63$) than adults ($M = 2.16$, $SD = 0.62$), $F(1, 124) = 3.05$, $p = .083$, partial $\eta^2 = .02$, but no other main effects or interactions approached significance. In contrast, in the engineering condition, there was a marginally significant effect in the reverse direction, such that those who correctly stated that weight

mattered were actually more confident (N = 54, M = 2.30, SD = 0.69) than were those who incorrectly stated that weight did not matter (N = 12, M = 1.89, SD = 0.58), $F(1, 124) = 3.42$, $p = .067$, partial $\eta^2 = .03$. For an analogous test with material, there were no significant effects or interactions, but there was a marginal effect of accuracy, such that those who correctly stated that material did not matter were more confident (N = 77, M = 2.22, SD = 0.66) than those who incorrectly stated that it does (N = 55, M = 1.93, SD = 0.77), $F(1, 124) = 2.77$, $p = .098$, partial $\eta^2 = .02$. Overall, it seems those who were more accurate, particularly in the science condition, were more likely to be confident in their answers.

Qualitative Explanations. The open-ended responses to questions of why the two levels of each variable are the same or different were coded for mention of data and of mechanism for both the science and engineering conditions. Mention of data was coded when participants explicitly mentioned numbers (e.g., "The short string went about 8.8 s and the long string was 12.4"), the number of times one setup went faster (e.g., "The heavy weight went faster most of the time"), or the relative difference between times (e.g., "The difference [between heavy/light] isn't significant compared with length"). Mention of mechanism was coded when participants offered a reason for the effect based on characteristics of the variable, such as, "The short string has less distance to go"; "Gravity pulls the heavy weight faster"; or "The wire doesn't bend." The codes were not mutually exclusive, so participants could be coded for mentioning both data and mechanism in one reply. Two researchers each coded all responses. In the science condition, there were 374 responses, yielding a 96.5% agreement level (Cohen's kappa = .930). In the engineering condition, they coded 609 responses, yielding a 98.4% agreement level (Cohen's kappa = 0.967). Discrepancies were resolved through discussion with the lead author.

The phase II test questions were not identical in each condition. In the science condition, after each variable was tested, participants were asked which level of the variable made the pendulum swing faster and why (or why they went the same speed, if that was their conclusion). In this condition, proportions of participants mentioning data as an explanation for beliefs about length (30%), weight (55%), and material (55%) were significantly different, as assessed by the Cochran test, which evaluates differences among related proportions, $\chi^2(2, N = 66) = 18.96$, $p < .001$. The Kendall coefficient of concordance was .14. Far more participants said data influenced their conclusions about weight and material than about length. The reverse pattern was found for mention of mechanism, where far more participants mentioned mechanism when explaining length (74%) than when explaining either weight (44%) or material (44%), as assessed by the Cochran test, $\chi^2(2, N = 66) = 23.14$, $p < .001$. The Kendall coefficient of concordance was .18. The patterns were similar across age groups.

In the engineering condition, participants were asked to explain which variables made the pendulum go fastest (or slowest) and why. In this condition, almost all participants justified their conclusions by talking about mechanisms (95.5%), while only a handful mentioned data (4.5%); two participants mentioned both, and one person mentioned neither. However, the frequency of mentioning mechanisms differed between the variables—a larger proportion of participants mentioned a mechanism for length (.94) than for weight (.63) or for material (.32). A Cochran test confirmed a significant difference, $\chi^2(2, N = 66) = 51.06$, $p < .001$, with a Kendall coefficient of concordance of .39. Follow-up pairwise comparisons using a McNemar's test showed all pairwise differences were significant at the $p < .01$ level.

Discussion

The results from the current study indicate that a systematic testing of variables, even without additional direct attention to the outcome data, is sufficient for nearly all adults and the majority of children to facilitate belief revision. When children and adults were guided to compare each level of a variable while holding everything else constant, they were able to learn that—in contrast to their initial beliefs—there was no effect of the noncausal variables. However, when given the engineering goal of designing a fast or slow pendulum, they had more difficulty in revising their beliefs about what, in fact, were noncausal variables.

The science condition required careful direct comparison of empirical results, the kind of controlled tests scientists ideally perform to learn about causal relationships. The task provided structure and guidance, while highlighting the most important characteristics as a feature of the design and making it straightforward to compare the role of each variable at different levels. In the science condition, participants were focused on the relevant mechanism information by isolating single variables and then by setting up the direct comparisons. This structure provided informative data that could help in drawing conclusions. In contrast, the engineering condition offered more latitude. There was still structure, and the task goal was experimenter-provided. However, there was freedom to choose how to set up the pendulum and, thus, to determine the precise types of evidence available to draw conclusions. In the science condition, participants had to interpret the evidence provided. In the engineering condition, they had the additional step of choosing which evidence to acquire, before interpreting the evidence, which required additional cognitive load.

Thus, guiding children to use systematic testing of variables by presenting them with a scientific goal rather than an engineering goal may be particularly effective when children are first learning about a task. Although explaining the logic of different goals is not very effective in leading to conceptual change (Siler et al., 2013), requiring a set goal may be effective, at least as a first step toward belief revision. Future research is needed to explore combinations of these two approaches—perhaps first requiring a scientific goal and then later explaining the logic with a concrete example may be effective for more long-term conceptual change. It is important to note that in our specified conditions, the comparisons were spelled out in the science condition and not in the engineering condition. Thus, the engineering condition required both designing comparisons and then interpreting the results. Future work comparing the scientific and engineering approaches with varying levels of constraints on the task will provide a more thorough understanding of the relative contributions of the scientific and engineering approaches as compared to the specificity of the tasks required.

The fact that children—and some adults—were not completely consistent in their replies to similar questions about each variable also suggests possible confusion about the questions. Children might have been confused by our asking a rephrased question when they had just addressed the issue; they might have presumed that the experimenter wanted a different answer.

The other major manipulation, the attention-to-data condition, did not lead to belief revision. The lack of demonstrated effect of this intervention indicates that the specific manipulation used may not have been strong enough in calling participants' attention to the data, particularly for the children. For example, the task that directed participants to compare pairs is more in line with a sign test than a complete comparison of lists, more closely approximating a *t* test comparison. It is also possible that children saw the data and their reasoning as not connected; if they were not

relying on the data in drawing their conclusions, even drawing their attention to the data may not have been helpful in aiding their application of the data to revising beliefs. The fact that the participants using an engineering goal rarely mentioned data as justification for their beliefs supports this interpretation. Kloos and Somerville (2001) suggested children may have difficulty with information consolidation; children may see evidence but not link it with prior knowledge or not know how to do so without guidance. Although not all participants mentioned data in justifying their reasoning in the science condition, the greater frequency of this response suggests the systematicity of testing is guiding which aspects of the evidence participants are considering. Thus, an important area for future study is to help children develop more explicit links between the empirical data they see and their conclusions.

GENERAL DISCUSSION

These two studies demonstrate that children and adults can use empirical data to revise their beliefs but that the process depends on the context in which the data are generated and interpreted and the goal that drives the exploration. In Study 1, most children and adults were able to update their incorrect beliefs after seeing systematic tests of each potential variable. They recognized the difference between the small variation of "error" and the larger variation between the long- and short-string conditions, which indicated a true underlying effect. However, in Study 2, when given the opportunity to explore the pendulum context with a specific engineering goal, children and adults had more difficulty using evidence to change their causal concepts to noncausal ones.

These findings demonstrate that the goal and testing approach in an experimental context are important in determining how children and adults learn from empirical explorations. In the current study, these goals were provided to participants in both the science and engineering conditions, and they affected learning. When children and adults generated their own testing approach to fulfill the provided goal, their resulting explorations led to less effective learning about causal mechanisms.

To have empirical data change one's beliefs, the data must be viewed as relevant and informative. Thus, it is necessary to understand how theory and data inform one another. When there is a clear explanation for a pattern of results, conceptual change is more likely (Ahn et al., 1995; Koslowski, 1996). Indeed, in the current study, when describing a finding that confirmed prior beliefs, children and adults were likely to offer explanations based on mechanism rather than on empirical data, whereas when describing a finding that disconfirmed prior beliefs, they were much less likely to invoke mechanism in their explanations. When the only added information is empirical data, the data need to be integrated with prior beliefs or discounted in some way (Chinn & Brewer, 1998). Further exploration of potential methods of drawing attention to data effectively for children and adults, perhaps with a more explicit lesson about the link between data and conceptual understanding, might lead to improved learning outcomes. Children may have gotten stuck at the observation or interpretation phase (Chinn & Malhotra, 2002) and misinterpreted what they saw to fit with prior beliefs, while adults were more effective at using the evidence. However, the structure of the guided scientific task made the observation process more clear, and that, in turn, made changes in the internal representation of the problem more likely for both children and adults, consistent with previous work supporting the value of direct instruction in many science education contexts (e.g., Klahr & Nigam, 2004).

The age differences that did occur (children struggling to learn more than adults, particularly on the more unstructured tasks) have several possible explanations. As with all cross-sectional studies, we have identified two snapshots of different age groups that suggest there is development across time. Future work with a microgenetic approach (cf. Siegler & Crowley, 1991) would provide us with a more nuanced picture of the development of this essential scientific skill. These cross-sectional data cannot directly address any particular interpretation, though the continuing difficulties with formal data analysis documented into adulthood (e.g., delMas & Liu, 2007; Konold & Pollatsek, 2002; Lubben et al., 2001; Volkwyn et al., 2008) suggest that the full understanding of these connections is likely not typically obtained even by age 12 years and thus requires more experience and cognitive skills.

Overall, these results support the hypothesis that a focus on numerical empirical data can lead to a change in understanding. Further work is needed to explore longer-term effects of these types of interventions and to determine the extent to which the knowledge gained would transfer both in domain knowledge regarding causal factors and in understanding the value of controlled testing of variables. There is also more to learn about the types of numerical empirical evidence that lead to belief revision. In the current studies, the two levels of the length variable did not have any overlapping values, yet many causal factors could yield more overlapping data. Here, our goal was to demonstrate that numerical data were sufficient to lead to belief revision and therefore opened the door for further explorations of the boundary conditions of these effects.

In sum, these studies demonstrate that systematically presented empirical evidence can effect belief revision. Given appropriate guidance, even fourth- and fifth-grade children can use empirical evidence to revise causal misconceptions, both to learn which factors are causal and, more impressively, to learn which are noncausal in a physical system. Moreover, they can revise flawed causal theories that are often held into adulthood. In contrast, when focused on an engineering goal (i.e., to achieve an effect), children were less able to utilize empirical evidence for belief revision—a finding that indicates it is not evidence of *any* kind that facilitates conceptual change, but rather the systematic evidence produced by controlled comparisons. Future work that examines how types of evidence may interact with goals in facilitating longer-term retention is needed to further our understanding of these issues. However, it is clear that numerical data, when presented in context with appropriate structure, can help children and adults reexamine their beliefs and support the process of conceptual change and robust scientific thinking.

ACKNOWLEDGMENTS

We thank Bryan Matlen, Stephanie Siler, and Brian Meagher for comments on an earlier draft. Thanks also to the students, principals, teachers, and staff at Winchester-Thurston School and Covert Elementary School for their participation. We also thank Audrey Russo, Mandy Jabbour, Stacey Gagnon, Ashley Porter, and Alyssa Dunn for assistance with data collection, data entry, and coding. Sadly, Mike Anzelone, one of the research assistants who helped tremendously with Experiment 2, died in August 2014. He was working toward his Ph.D. in Clinical Psychology, and it is a loss both personally and for the field.

FUNDING

This work was supported in part by a grant from the National Science Foundation (BCS-0132315) to David Klahr.

REFERENCES

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352. doi:10.1016/0010-0277(94)00640-7

Allen, M. (2010). Learner error, affectual stimulation, and conceptual change. *Journal of Research in Science Teaching*, *47*, 151–173.

Amsel, E., Goodman, G., Savoie, D., & Clark, M. (1996). The development of reasoning about causal and non-causal influences on levers. *Child Development*, *67*, 1624–1646. doi:10.2307/1131722

Bonawitz, E. B., van Schijndel, T. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*, 215–234. doi:10.1016/j.cogpsych.2011.12.002

Burbules, N. C., & Linn, M. C. (1988). Response to contradiction: Scientific reasoning during adolescence. *Journal of Educational Psychology*, *80*, 67–75. doi:10.1037/0022-0663.80.1.67

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*, 1098–1120. doi:10.1111/1467-8624.00081

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405. doi:10.1037/0033-295X.104.2.367

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. doi:10.1002/(ISSN)1098-2736

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction*, *19*, 323–393. doi:10.1207/S1532690XCI1903_3

Chinn, C. A., & Malhotra, B. A. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology*, *94*, 327–343. doi:10.1037/0022-0663.94.2.327

delMas, R., & Liu, Y. (2007). Students' conceptual understanding of the standard deviation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data (Proceedings of the 33rd Carnegie Symposium on Cognition)* (pp. 87–116). New York, NY: Erlbaum.

Frick, A., Huber, S., Reips, U., & Krist, H. (2005). Task-specific knowledge of the law of pendulum motion in children and adults. *Swiss Journal of Psychology/Schweizerische Zeitschrift Für Psychologie/Revue Suisse De Psychologie*, *64*, 103–114. doi:10.1024/1421-0185.64.2.103

Fugelsang, J. A., & Thompson, V. A. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition*, *31*, 800–815. doi:10.3758/BF03196118

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence* (A. Parsons & S. Milgram, Trans.). New York, NY: Basic Books. (Original work published 1955)

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*, 748–769. doi:10.1002/(ISSN)1098-2736

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*, 75–86. doi:10.1207/s15326985ep4102_1

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*, 661–667. doi:10.1111/j.0956-7976.2004.00737.x

Kloos, H., & Somerville, S. C. (2001). Providing impetus for conceptual change: The effect of organizing the input. *Cognitive Development*, *16*, 737–759. doi:10.1016/S0885-2014(01)00053-3

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*, 259–289. doi:10.2307/749741

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

Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). The development of scientific thinking skills. In *Developmental psychology series*. Orlando, FL: Academic Press.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal*, *41*, 635–679. doi:10.3102/00028312041003635

Limón, M. (2002). Conceptual change in history. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 259–289). Dordrecht, The Netherlands: Kluwer Academic.

Limón, M., & Carretero, M. (1997). Conceptual change and anomalous data: A case study in the domain of natural sciences. *European Journal of Psychology of Education*, *12*, 213–230. doi:10.1007/BF03173085

Lovett, M. C., & Chang, N. M. (2007). Data-analysis skills: What and how are students learning? In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 293–318). New York, NY: Taylor & Francis.

Lubben, F., Campbell, B., Buffler, A., & Allie, S. (2001). Point and set reasoning in practical science measurement by entering university freshmen. *Science Education*, *85*, 311–327. doi:10.1002/sce.1012

Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education*, *18*, 955–968. doi:10.1080/0950069960180807

Masnick, A. M., & Klahr, D. (2003). Error matters: An initial exploration of elementary school children's understanding of experimental error. *Journal of Cognition and Development*, *4*, 67–98. doi:10.1080/15248372.2003.9669683

Masnick, A. M., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development*, *79*, 1032–1048. doi:10.1111/cdev.2008.79.issue-4

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Research Council (NRC). (1996). National science education standards. In *National committee on science education standards and assessment*. Washington, DC: National Academy Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in Grades K–8*. Washington, DC: National Academies Press.

Next Generation Science Standards (NGSS) Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press.

Penner, D. E., & Klahr, D. (1996). When to trust the data: Further investigations of system error in a scientific reasoning task. *Memory & Cognition*, *24*, 655–668. doi:10.3758/BF03201090

Petrosino, A., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning*, *5*, 131–156. doi:10.1080/10986065.2003.9679997

Schauble, L., Klopfer, L., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, *28*, 859–882. doi:10.1002/(ISSN)1098-2736

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*, 1045–1050. doi:10.1037/0012-1649.43.4.1045

Schwartz, D. L., Sears, D., & Chang, J. (2007). Reconsidering prior knowledge. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 319–344). New York, NY: Taylor & Francis.

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, *46*, 606–620. doi:10.1037/0003-066X.46.6.606

Siler, S. A., & Klahr, D. (2012). Detecting, classifying and remediating children's explicit and implicit misconceptions about experimental design. In R. W. Proctor & E. J. Capaldi (Eds.), *Psychology of science: Implicit and explicit reasoning*. New York, NY: Oxford University Press.

Siler, S., Klahr, D., & Matlen, B. (2013). Conceptual change in experimental design: From engineering goal to science goals. In S. Vosniadau (Ed.), *Handbook of research on conceptual change* (2nd ed., pp. 138–158). New York: Routledge.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318. doi:10.1016/j.tics.2006.05.009

Volkwyn, T. S., Allie, S., Buffler, A., & Lubben, F. (2008). Impact of a conventional introductory laboratory course on the understanding of measurement. *Physical Review Special Topics: Physics Education Research*, *4*, 010108-1–010108-10. doi:10.1103/PhysRevSTPER.4.010108

Vosniadou, S. (2007). The conceptual change approach and its re-framing. In S. Vosniadou, A. Baltes, & X. Vamvakoussi (Eds.), *Re-framing the conceptual change approach in learning and instruction* (pp. 1–15). Amsterdam, The Netherlands: Elsevier.

Vosniadou, S. (2008). *International handbook of research on conceptual change*. New York, NY: Routledge, Taylor & Francis.

Vosniadou, S. (2013). *International handbook of research on conceptual change* (2nd ed.). New York, NY: Routledge, Taylor & Francis.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*, 535–585. doi:10.1016/0010-0285(92)90018-W

White, P. A. (2003). Causal judgement as evaluation of evidence: The use of confirmatory and disconfirmatory information. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *56*, 491–513. doi:10.1080/02724980244000503

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*, 172–223. doi:10.1016/j.dr.2006.12.001