Child Development, July/August 2008, Volume 79, Number 4, Pages 1032 – 1048

Investigating the Development of Data Evaluation: The Role of Data Characteristics

Amy M. Masnick
Hofstra University

Bradley J. Morris
Grand Valley State University

A crucial skill in scientific and everyday reasoning is the ability to interpret data. The present study examined how data features influence data interpretation. In Experiment 1, one hundred and thirty-three 9-year-olds, 12-year-olds, and college students (mean age = 20 years) were shown a series of data sets that varied in the number of observations and the amount of variance between and within observations. Only limited context for the data was provided. In Experiment 2, similar data sets were presented to 101 participants from the same age groups incrementally rather than simultaneously. The results demonstrated that data characteristics affect how children interpret observations, with significant age-related increases in detecting multiple data characteristics, in using them in combination, and in explicit verbal descriptions of data interpretations.

One crucial skill in scientific and everyday reasoning is the ability to interpret data. For example, imagine two golfers on a driving range hit three balls each. Golfer A hits the balls 100, 130, and 125 yards, whereas Golfer B hits the balls 250, 265, and 270 yards. If asked to predict which golfer would likely hit the ball farther on the next drive, an observer would probably pick Golfer B. It is likely that our observer would not treat each drive as an unrelated incident but would see each drive as one piece of data related to a stable underlying construct (e.g., skill in driving a golf ball). Along with this understanding is an expectation that even when such a construct is stable, there is variation in outcomes (i.e., exact distance of drive). Although the mechanism causing the difference might not be clear to the observer (e.g., strength, practice), the resulting pattern of observations suggests nontrivial differences between the performances of the two golfers. Explanations for the difference in performance (i.e., specifying causal mechanism) would arise from a combination of the observer's background knowledge and the data.

Although such inferences may be trivial for adults, children often err when making similar comparisons (see Zimmerman, 2000, 2007, for reviews). One source of difficulty is that children tend to rely heavily on their domain knowledge when interpreting data, even to the extent of ignoring disconfirming data that conflict with their current knowledge or expectations (e.g., Chinn & Malhotra, 2002; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Schauble, 1996). For example, when children observed the speed of two falling objects that differed in size or weight, their prior beliefs that the objects would fall at different speeds affected both their observations and conclusions from the data. Thus, after viewing two objects released at the same time from the same height, many children claimed that they had observed one object fall faster than the other, despite the fact that they fell at the same rate (Chinn & Malhotra, 2002).

Many investigations of children's data interpretation have done so within knowledge-rich domains (e.g., Echevarria, 2003; Metz, 2004). However, children often encounter situations about which they have very limited background knowledge. In these

This work was supported in part by grants from the National Institute of Child Health and Human Development (HD25211) to David Klahr, and from the National Institute of Mental Health training Grant T32 MH19102. Thank you to the parents, teachers, and principals at St. Teresa of Avila, Sterrett Classical Academy, Greenfield Elementary, Sacred Heart Elementary, Schiller Classical Academy, Colfax Elementary, and the Frick Academy for their invaluable assistance. We also owe many thanks to Anne Siegel, Audrey Russo, and Jen Schnakenberg for assistance with data collection and coding; to Steve Tibbets and Julie Heath for assistance with SAS programming; and to David Klahr, Corinne Zimmerman, and three anonymous reviewers for helpful comments on earlier drafts. Portions of the preliminary results from Experiment 1 were summarized in Masnick, A. M., Klahr, D., & Morris, B. J. (2007). Separating signal from noise: Children's understanding of error and variability in experimental outcomes. In M. Lovett & P. Shah (Eds.) *Thinking with data* (pp. 3 – 26). New York: Lawrence Erlbaum. In addition, portions of these results were reported at the meetings of the American Educational Research Association (2002 and 2004), the Society for Research in Child Development (2003 and 2005), and the Cognitive Science Society (2002 and 2004).

Correspondence concerning this article should be addressed to Amy M. Masnick, Department of Psychology, Hofstra University, Hempstead, NY 11549. Electronic mail may be sent to amy.m.masnick@hofstra.edu.

situations, data provide a rich source of information from which to draw conclusions about a phenomenon. What is not known is whether children attend to and use particular diagnostic features of data. That is, given a set of observations, are there properties of data sets—unrelated to knowledge of the domain—that help reasoners induce trends from data sets without using formal statistics?

We suggest that two diagnostic features of data patterns may play an important role in reasoning: sample size and variability. Sample size refers to the number of observations within a data set. Considering the golfing example mentioned previously, someone using sample size as a factor to assess differences between two golfers would likely be more confident about drawing conclusions with more data available (e.g., three hits per golfer vs. one hit per golfer). Variability refers to the degree of similarity or dissimilarity among values within and between the set of observations. Returning to the golfing example, *within-group* variability is a measure of differences in the outcomes of drives made by the same golfer, whereas *between-group* variability is a measure of differences in drives between golfers. A reasoner using variability characteristics of data would be more confident of conclusions drawn from data as the variability within a set of observations decreases and variability between a set of observations increases.

Reasoning About Data in Category Induction and Scientific Reasoning Tasks

In category membership tasks, there is evidence that elementary school children recognize sample size and variability. Gutheil and Gelman (1997) found that although adults used variability and sample size in combination or independently, 8- and 9-year-old children did not differentiate between sets of exemplars that differed only in sample size or in variability but did differentiate when both sample size and variability differed between sets. In related work, Jacobs and Narloch (2001) asked children in Grades 1, 3, and 5 to predict how common a trait would be in a given population based on a sample that varied in number (sample size) and homogeneity (within-group variability). Prior knowledge about variability affected children's predictions more than sample size. Strong prior assumptions about the phenomena in question influence reasoning about data characteristics.

There is also evidence that children attend to sample size and data variability in scientific reasoning contexts. Piaget and Inhelder (1951/1975) investigated children's assessments of the probability of a spinner landing in a series of possible locations. By middle childhood, children's judgments about future landing sites were related to the number of previous observations, indicating some understanding of sample size (i.e., the law of large numbers, or the idea that larger sample sizes make one more confident about generalizing to a population). Klaczynski and Aneja (2002) also demonstrated that children as young as 7 years old applied the law of large numbers in inferring new gender stereotypes by being more confident when shown a larger sample than a small one.

A very common scientific reasoning task involves drawing conclusion based on the evaluation of covariation data (e.g., Koerber, Sodian, Thoermer, & Nett, 2005; Kuhn, Amsel, & O'Loughlin, 1988; Shaklee & Paszek, 1985). Interpreting covariation data is also influenced by data characteristics. Koslowski, Okagaki, Lorenz, and Umbach (1989) asked sixth-grade, ninth-grade, and college participants whether a target factor was likely to cause a particular effect (e.g., whether using a gasoline additive leads to worse gas mileage). Information presented to participants varied in sample size and in whether there was evidence of covariation. Without covariation information, all participants differentiated between large and small sample sizes; however, with covariation information, only the college students used sample size information. Thus, there is evidence that sixth and ninth graders have some difficulty integrating multiple characteristics of the data. At the same time, even sixth graders can use sample size in the absence of other information to infer cause. Koslowski (1996, Experiment 10) also reported that when more instances of covariation were present, participants were more likely to rate the association between two variables as causal than when there was only one instance of covariation.

Children also attend to variability within and between data sets. When second and fourth graders rolled two balls down two ramps and measured the distance each traveled, they distinguished between the small differences expected in individual data points (within-group variability) and larger differences expected between groups (between-group variability; Masnick & Klahr, 2003). The expectation that there will be some variation in precise data points indicates an understanding that many factors affect the outcome of an experiment, even when these small differences may not affect overall conclusions.

Yet, there is evidence that children have difficulty understanding what this variability indicates. Lubben and Millar (1996) showed 11-, 13-, and 15-year-olds two sets of results with equal means but different amounts of variation. Participants were asked whether one set of results was more trustworthy or if they were equally trustworthy. About half (47% – 55%) of all participants

said that one data set was more trustworthy than the other. However, only about 20% of 11-year-olds justified their choice by referring to variation in the data. Forty-four percent of 13-year-olds and 48% of 15-year-olds justified their choice by referring to variation in data, indicating an understanding that data variability is a component of assessing reliability (Lubben & Millar, 1996). Thus, even into high school, a sizable number of students have difficulty using variability information in drawing conclusions.

Studies of covariation detection, in effect, ask participants to assess an ''intuitive chi-square'' estimate: How likely is it that the observed pattern of frequencies occurred by chance? Although much of the research on the evaluation of evidence has taken the form of covariation matrices concerning the frequency of the presence and absence of putative causes and effects (i.e., an intuitive chi-square, as in Shaklee & Paszek, 1985), in both real-world contexts and the science classroom, children must evaluate the differences in a range of *quantitative* data. Zimmerman (2000) noted that to determine if a specific antecedent is linked with an outcome, one compares the number of times the outcome occurs or does not occur when the antecedent is present. If the former is much larger than the latter, one can generally conclude that there is a relationship between the two. However, she observed that ''it is not clear *how large* the difference must be in order to conclude that the two events are related'' (p. 115). The question is then, How do students reason when comparing sets of quantitative data?

In addition to the evidence that elementary and secondary school children attend to features of data in some contexts, there is also evidence that very young children have some fundamental understanding of relative quantity. Even 6-month-old infants can accurately compare two quantities (see Feigenson, Dehaene, & Spelke, 2004, for a discussion), and 5-year-old children are equally adept at determining which quantity is larger when presented Arabic numerals or arrays of dots (Temple & Posner, 1998). Thus, it seems possible that some of the skills required for effective data comparison are implicit and not the result of direct training. It is possible that children use a relatively automatic process of quantity comparison. We follow Rubinsten, Henik, Berger, and Shahar-Shalev (2002) in suggesting that ''automatic'' refers to processing that requires no conscious monitoring after commencing. The process that operates may be an ''intuitive'' statistical test on data patterns. We use the term ''intuitive'' to convey the automaticity of the process.

Clearly, comparing sets of data requires processing beyond that of simply detecting differences in magnitude. A second set of processes is needed to process exact differences (e.g., perform calculations), to examine characteristics of data, and to integrate this information with domain knowledge. Inhelder and Piaget (1958) suggested that a pattern of fixating on one feature instead of looking at features in concert was the source of young children's difficulty in solving conservation tasks. If children do focus initially on a small number of features, then we would expect to see age-related increases in the number of characteristics to which they attend. Additionally, children would likely develop the ability to compare sets along multiple dimensions before they could explicitly access knowledge to consider each variable independently within the set. It is likely that these improvements are due to many factors including acquisition of new strategies, increases in domain knowledge, and increases in processing capacity (Halford, Cowan, & Andrews, 2007).

Present Study

To our knowledge, there has been no systematic investigation of the characteristics of data to which children attend and to what extent these characteristics influence their judgments about phenomena. The present study was designed to examine which characteristics of data guide inferences when comparing data sets. First, we predicted that sample size is likely to play an important role in reasoning. Second, we predicted that in drawing conclusions about comparative data, two characteristics that indicate the *amount of variation* in the data are key: within-group variability (i.e., the variability of the data points relative to mean) and between-group variability (i.e., variability of data points in each of two data sets relative to one another). This information about variability can be assessed increasingly well with a larger sample size of data points. Finally, we predicted that with age, children will become better at using multiple characteristics of data in concert and will have more explicit understanding of the importance of these characteristics for making judgments about quantitative data.

We chose to work with third- and sixth-grade students to explore developmental changes through elementary school because past work on reasoning with data has often examined children in these age groups (e.g., Gutheil & Gelman, 1997; Inhelder & Piaget, 1958; Jacobs & Narloch, 2001; Koslowski et al., 1989), as they become exposed to experimental data in school. Our pilot data indicated that third graders were the youngest age group in which all participants could consistently perform our data interpretation

tasks. A 3-year age difference allows for a large enough difference to see change if it is apparent. We chose a college student comparison population because it is clear that these skills continue to develop with experience and age.

In two experiments, children and college students were presented with sets of paired data and asked to draw conclusions about differences between the sets. In Experiment 1, data were presented with one of two framing stories that limited the amount of background knowledge that could be brought to bear in drawing conclusions about the data. The data presented varied systematically in number of data points presented, within-group variability (operationalized as a dichotomous variable, with two sizes of standard deviations relative to the mean), and between-group variability (operationalized as the number of pairs of data in which the data point from the column with the lower mean was higher than the data point from the column with the higher mean).

When studying reasoning about any topic, it is important to consider both implicit and explicit evidence of knowledge used. Nisbett and Wilson (1977) argued that explicit justifications for behaviors are often distinct from actual reasons for behavior and that the salience of a stimulus is the key determinant of it being used as a justification. Thus, a full understanding of the cognitive process involves looking at both patterns of conclusions drawn from data and explicit rationales for these patterns. Participant responses were analyzed to examine how each of the three characteristics of data (sample size, between-group variability, and within-group variability) was used in implicit and explicit reasoning.

Experiment 1

Method

Participants. Thirty-nine third graders (mean age = 9.1 years, range = 8.2–10.3 years), 44 sixth graders (mean age = 11.9 years, range = 11.2–12.8 years), and 50 college undergraduates (mean age = 20.2 years, range = 18.1–23.7 years) participated in this study. College students were recruited from undergraduate psychology courses, and younger participants were recruited from letters sent to parents at four elementary and middle schools in the northeast United States. The third-grade sample consisted of 92% White students, 5% Black students, and 3% Hispanic students. The sixth-grade sample consisted of 82% White students, 2% Asian students, and 16% Black students. The college student sample consisted of 54% White students, 40% Asian students, 2% Black students, and 4% Hispanic students.

Procedure. All participants were interviewed individually. Participants were randomly assigned to one of two cover story conditions: one in which robots were the source of the data and one in which athletes (people) were the source of the data. In the robot condition, each participant was read the following information:

> Some engineers are testing new sports equipment. Right now, they are looking at the quality of different sports balls, like tennis balls, golf balls and baseballs. For example, when they want to find out about golf balls, they use a special robot launcher to test two balls from the same factory. They use a robot launcher because they can program the robot to launch the ball with the same amount of force each time. Sometimes they test the balls more than once. After they run the tests, they look at the results to see what they can learn.

In the athlete condition, we used an isomorphic cover story in which two athletes were trying out for one slot on a sports team. The coaches asked the participants to perform certain tasks (e.g., hit a golf ball as far as possible) to assess which athlete would be better for the team. For example, participants in the athlete condition saw the following story: ''Harriet and Joan are trying out for the soccer team. The soccer coach asked them to kick a soccer ball four times. The coach measured how far the soccer ball went each time it was kicked.'' This cover story was designed to see if adding information about a potential source of variability (human error) would change participants' responses. Thus, the robot condition presented a cover story that minimized potential sources of variability, whereas the athlete condition provided potential sources of variability.

After reading the cover story, the participants were shown a series of 14 data sets, one at a time. For each data set, there were data either for two balls of the same type, with no distinguishing characteristics (e.g., Baseball A and Baseball B), or for two athletes, with no information other than first names (e.g., Alan and Bill). The 14 balls tested were golf balls, racquetballs, basketballs, soccer balls, ping pong balls, footballs, volleyballs, baseballs, tennis balls, marbles, hockey pucks, kickballs, softballs, and bowling balls. Participants were randomly assigned to one of two orders of presentation. There were no order effects; therefore, results are collapsed across order. In the athlete condition, different names were used for each data set to prevent any carryover knowledge effect.

For each data set, there were one, two, four, or six pairs of data. Each page contained two columns of data: one column listing the distance the first ball traveled and one listing the distance the second ball traveled. (See Table 1 for specific examples of the different data characteristics.)

The 14 data sets varied in (a) sample size, (b) between-group variability (i.e., reversals in which column's number was higher), and (c) within-group variability (i.e., high or low variability relative to the means). Each participant evaluated 14 comparisons, with eight trials including no reversed pairs (sample sizes 1, 2, 4, and 6) and six trials including one or two reversed pairs (sample size 4 with one reversed pair, and sample size 6 with one and two reversed pairs). Half of the trials had high within-group variability, in which the standard deviation of the data set was 15% – 20% of its mean, and half had low within-group variability, in which the standard deviation of the data set was less than 2% of its mean. Each of the 14 trials tested a different type of sports ball.

For each data set, participants were asked first what the engineer or coach could find out as a result of this information and to explain any reasons for their answer. Next, they were asked how sure they were about these conclusions. To answer the questions about sureness, participants were offered a 4-point scale from which to select their answer, choosing among *not so sure*, *kind of sure*, *pretty sure*, and *totally sure*.

Measures. For each question in which participants were asked to report their conclusions and predictions, and their confidence in their answers, participants first answered yes or no and then rated their confidence on a 4-point Likert scale. These two responses—yes – no and sureness level—were combined into a single 7-point ordinal variable: *totally sure there is a difference*, *pretty sure there is a difference*, *kind of sure there is a difference*, *not so sure* (regardless of yes or no answer), *kind of sure there is no difference*, *pretty sure there is no difference*, *totally sure there is no difference*. Past research has demonstrated that the children as young as 8 years old have been able to understand and work with such a scale (Masnick & Klahr, 2003). Participants were asked to provide reasons for their initial conclusions and final predictions of relative position. These reasons were coded for mention of data characteristics (such as sample size or variability) or mechanism (such as a property of the ball that could have affected the results).

Table 1
Examples of Data Sets in Experiment 1

Example 1: Six data pairs with one reversed data pair (five of six times Basketball B goes farther), high within-group variability, robot condition

Basketball A	Basketball B
56 ft	74 ft
65 ft	83 ft
75 ft	57 ft
49 ft	66 ft
48 ft	67 ft
64 ft	82 ft

Example 2: Two data pairs with no reversed pairs, low within-group variability, athlete condition

Joan	Harriet
372 ft	383 ft
363 ft	374 ft

Results

The main research question was whether different types of data patterns and different contexts affect participants' decisions about, and confidence in, their judgments of whether there is a difference between the two groups of data presented. Decisions about whether the data sets differed were operationalized by participants' responses to whether they thought the engineer or coach could be sure that there was a difference between the two groups. In addition, participants' reasons for the responses were analyzed.

Use of data characteristics to judge whether there was a difference between the data sets. For each assessment, sample size, between-group variability (number of reversed pairs), and within-group variability (spread) were within-subjects variables, and cover story was a between-subjects variable. There were no gender differences. In addition, for the college students, there were no differences in ratings for those who had taken different numbers of statistics classes. Thus, these two variables were not considered in later analyses. Because the data were not set up in a full factorial design, each data characteristic was analyzed separately. Of those who said that one ball traveled farther than the other, third graders were accurate in stating which ball went farther 92.4% of the time, sixth graders were accurate 98.8% of the time, and college students were accurate 99.5% of the time.

A 4 (sample size: 1, 2, 4, or 6) $\times$ 3 (age: third grade, sixth grade, or college) $\times$ 2 (cover story: robot vs. athlete) mixed analysis of variance (ANOVA) was used to assess whether sample size influenced confidence judgments in cases with no between-group variability. There was a main effect of sample size, $F(3, 120) = 20.68$, $p < .001$, partial $\eta^2 = .34$; a main effect of

age, $F(2, 122) = 44.52$, $p < .001$, partial $\eta^2 = .42$; no main effect of cover story, $F(1, 122) = 0.70$, $p > .1$; and a Sample Size × Age interaction, $F(6, 240) = 17.63$, $p < .001$, partial $\eta^2 = .31$ (Figure 1). Tukey's honestly significant difference (HSD) post hoc tests indicated that college students' ratings were significantly lower on average than the third and sixth graders'. College students' sureness ratings increased dramatically with sample size, $F(3, 44) = 50.59$, $p < .001$, partial $\eta^2 = .78$, and sixth graders' ratings increased a small but significant amount with increased sample size, $F(3, 38) = 3.33$, $p = .030$, partial $\eta^2 = .21$. Third graders, in contrast, showed a slight trend of *decreased* sureness ratings with increased sample size, $F(3, 34) = 2.26$, $p = .099$, partial $\eta^2 = .17$. There were no interactions with the cover story. In sum, at all age levels, it was apparent that participants were sensitive to differences in sample size. For sixth graders and college students, this observation tended to increase ratings of sureness; for third graders, it tended to decrease ratings.

In addition to sensitivity to sample size, we also explored sensitivity to the presence of reversed data pairs (between-group variability) by examining only the data sets with six pairs of data. This pattern occurred when there were no reversed data pairs, one reversed pair, and two reversed pairs. A 3 (between-group variability: zero, one, or two pair reversals) × 3 (age: third grade, sixth grade, or college) × 2 (cover story: robot vs. athlete) mixed ANOVA demonstrated no effect of cover story, $F(1, 123) = 0.56$, $p > .1$. There was a main effect for between-group variability (lower sureness ratings with more variability), $F(2, 122) = 53.15$, $p < .001$, partial $\eta^2 = .47$; a main effect for age (higher average ratings for third graders than for sixth graders and college students), $F(2, 123) = 9.59$, $p < .001$, partial $\eta^2 = .14$; and a Variability × Age interaction, $F(4, 244) = 4.70$, $p = .001$, partial $\eta^2 = .07$ (Figure 2). At all age levels, participants' mean sureness ratings decreased with more reversed data pairs between the two groups of data, with the effect stronger with increasing age: for third graders, $F(2, 34) = 3.51$, $p = .041$, partial $\eta^2 = .17$; for sixth graders, $F(2, 40) = 14.42$, $p < .001$, partial $\eta^2 = .42$; and for college students, $F(2, 46) = 68.15$, $p < .001$, partial $\eta^2 = .75$.

We also assessed differences in within-group variability (data spread) on ratings of confidence that there was a difference between the two data sets. Participants saw two data sets with each sample size and between-group variability combination used (e.g., two data pairs with no reversed pairs; six data pairs with one reversed pair). One of these data sets had high within-group variability (standard deviation between 15% and 20% of mean) and one had low within-group variability (standard deviation less than 2% of mean). We calculated a paired t test for each sample size/between-group variability combination participants saw (directly comparing sureness ratings from data with high and low within-group variability). Third and sixth graders showed no evidence of differentially rating data sets based on within-group variability. College students had an inconsistent pattern, in which three of the six comparisons reached significance (for four pairs with no reversed pairs, for six pairs with one reversed pair, and for six pairs with two reversed pairs). Surprisingly, in each of these comparisons, participants were more sure in the case with *higher* variability. Possible explanations for this pattern are proposed in the Experiment 1 Discussion section.

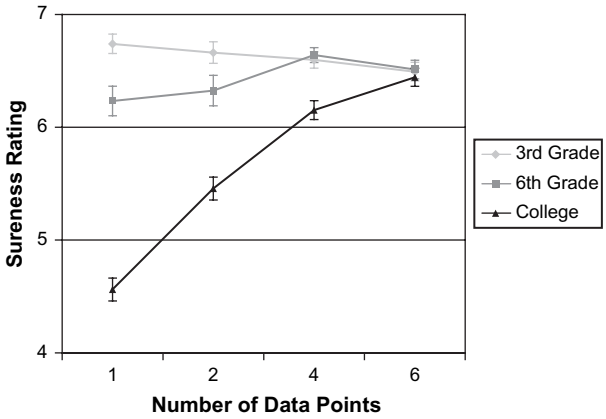

Figure 1. Experiment 1: Participants' sureness ratings at each grade with one, two, four, and six pairs of data but no reversed data pairs (collapsed across cover story).
Note. Error bars represent standard error of measurement.

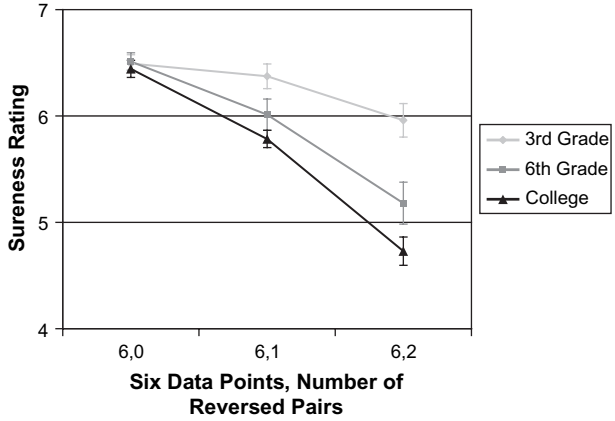

Figure 2. Experiment 1: Participants' sureness ratings at each grade with six pairs of data, by level of between-group variability in the data (collapsed across cover story).
Note. Error bars represent standard error of measurement.

When people evaluate differences between sets of data, they do not focus on only one characteristic. For example, in computing formal statistics, a *t* statistic takes account of both sample size and variability when testing for significant differences between means. The next analysis explored the possibility that our participants used an ''intuitive *t* test'' in which a combination of the number of data points and the variability of the data points were both considered in evaluating differences between means. Although these are small data sets, we calculated the independent *t*-test statistic for each set of data presented to the participants to examine the trends (with the obvious exception of the two sets with one pair). These 12 *t* values ranged from 0.88 to 4.18. Next, we correlated the *t* statistics for the presented data with the mean sureness rating participants gave assessing whether there was a difference between the two sets of data, by grade. An interesting pattern emerged. Third graders' ratings were not significantly associated with the independent *t* statistic ($r = .388$, $p = .213$), sixth graders' ratings showed evidence of a strong relationship ($r = .632$, $p = .028$), and college students showed evidence of a very strong relationship ($r = .873$, $p < .001$).

Reasons offered. Participants offered explicit justifications for confidence or lack thereof in their conclusions and for the predictions they made about which ball would go farther on a subsequent trial. Reasons given were coded for mention of the manipulated data characteristics: (a) sample size, (b) between-group variability (whether the data sets included pair reversals or not), and (c) within-group variability (spread of the data). We also noted mention of (d) a trend in the data and (e) the magnitude of the difference between the two data sets. Additionally, explanations of how the outcome could have occurred were coded, including (f) a property of the ball that affected the results, (g) a property of the robot or athlete that affected the results, or (h) a property of the environment—other than the ball, robot, or athlete—that affected the results. The categories were not mutually exclusive, and participants had 28 opportunities to explain their reasoning (twice for each of the 14 data sets). Participants' responses were coded by two independent coders, whose agreement level was 85%. All discrepancies were resolved via discussion.

Table 2 presents a summary of these results. An overwhelming majority of the responses were in reference to the data and not to theoretical issues such as properties of the ball or robot/athlete. Many participants mentioned the specific data characteristics that were manipulated, including sample size (e.g., ''They only tested it two times so they can't really be sure''), between-group variability characteristics such as reversed pairs (''All but once A went farther than B'') or no reversed pairs (''A went farther than B every time''), and within-group variability characteristics such as variability within the column (''All the numbers in A are close together'').

Table 2
Experiment 1: Percentage of Participants at Each Age Who Mentioned Each Justification for Their Sureness

	Third grade	Sixth grade	College
Data responses			
Sample size**	10	39	100
Reversed pairs*	56	70	90
No reversed pairs**	8	14	62
Spread of data**	3	15	44
Overall trend in data	92	93	100
Single pair ''trend''**	23	0	0
Magnitude of differences**	46	89	92
Magnitude of single pairs	5	2	0
Mechanism responses			
Ball property	15	20	18
Robot/athlete property	41	37	30
Environment	3	2	2

Note. Age differences: *$p < .05$. **$p < .01$.

In addition, other characteristics not manipulated were also mentioned. The most frequently mentioned data characteristic was a trend in the data (e.g., ''Five out of six times A went farther''; ''B went farther most of the time''). Some third graders (23%) explicitly referred to only one pair of the data set, effectively ignoring the remaining data in their justifications; we called this type of response a ''single-pair trend.'' No sixth graders or college students justified their reasoning this way. Another data characteristic mentioned was the magnitude of differences between the two columns of data (e.g., ''A went much farther than B''). In general, college students used a much wider range of responses than younger children, with nearly all of them mentioning sample size at least once. However, all participants, with the exception of 2 third graders, mentioned at least one data characteristic at some point during the study.

In addition to discussions of data, we also explored whether participants offered causal or mechanistic explanations as justifications for their conclusions (Table 2). We coded explanations for a mention of characteristics of the robot (e.g., ''Maybe the robot was breaking down after it threw Ball A''), characteristics of the athlete (e.g., ''Maybe Person A was having a good day''), characteristics of the ball (e.g., ''Maybe Ball B was more aerodynamic''), and characteristics of

the environment (e.g., ''Maybe there was wind blowing when Ball A was being thrown''). Across all trials, about half of the participants mentioned at least one mechanistic explanation, with no age differences (51% of third graders, 56% of sixth graders, and 42% of college students mentioned such an explanation).

One area in which a cover story difference might be expected is in use of these mechanistic justifications. Indeed, use of these mechanistic justifications did vary considerably by condition: Sixty-two percent of participants in the athlete condition said that a property of the athlete was a reason for at least one outcome, whereas 10% of those in the robot condition suggested a property of the robot as a reason for an outcome at least one time, $\chi^2(1, N = 133) = 38.19$, $p < .001$. In contrast, 6% of participants in the athlete condition and 29% of participants in the robot condition suggested a property of the ball as a potential explanation for the outcome at least one time, $\chi^2(1, N = 133) = 12.16$, $p < .001$. Mentions of other environmental mechanisms such as wind only occurred three times, twice in the robot condition and once in the athlete condition.

Discussion

When sample size, between-group variability, and within-group variability were manipulated, we found that sample size and the between-group variability influenced participants' reasoning but that within-group variability had no clear effect. In addition, older participants were able to make more accurate distinctions between data sets, indicating better knowledge of data characteristics individually and in concert. Children were also able to talk about data characteristics as early as in the third grade, demonstrating an early recognition of the importance of these concepts, but there were significant age-related increases in the number of data characteristics described in their explanations.

Although prior research indicates sensitivity to the law of large numbers at as early as age 7 (e.g., Klaczynski & Aneja, 2002), only college students indicated a clear appreciation of the law of large numbers by increased confidence ratings with increased sample size. Sixth graders showed a mixed pattern, and a sizeable number of third graders were less sure about conclusions when there were more data points—though this result does suggest sensitivity to the number of data points presented. These students may still be looking for a single correct answer and find multiple data points to be more confusing than informative.

In addition, there is evidence that in all three age groups, children pay attention to between-group variation in the data, one of the more complicated statistical concepts. Between-group variability in data is clearly salient to children even when the data are not couched in a strong theoretical context. Though this concept is relatively difficult in formal statistics, it may be implicitly provided by the presentation format of the data itself. For example, when scanning numbers presented in lists, when all numbers are visible simultaneously, and recall is not being tested, children may encode information about variability while encoding the values themselves. Thus, with age-related increases in processing capacity and speed as well as more optimal strategy use (e.g., Gathercole, Pickering, Ambridge, & Wearing, 2004; Kail, 2007), participants showed improvement in their ability to use information about sample size and between-group variability and to use the characteristics together in the form of an intuitive *t* test.

Participants did not seem to respond to changes in within-group variability, our manipulation of the standard deviation relative to the mean, though college students showed a small trend toward being more sure in cases with greater variability. It is possible that this characteristic was simply not salient or believed to be important and thus did not influence most responses. It is also possible that the lack of effect for children, and small reverse effect for college students, occurred because variability was confounded with data magnitude, such that the data with low variability relative to the mean consisted entirely of three-digit numbers, whereas the data with high variability relative to the mean consisted entirely of two-digit numbers. This confounding may have led to college participants simply rating themselves as more confident when reasoning about two-digit numbers instead of the more challenging three-digit numbers. In addition, with small sample sizes, the spread of the data may have been less noticeable.

One of our more intriguing findings was that participants appear to become increasingly adept with age at performing intuitive *t* tests on the data sets, that is, in using the data characteristics in concert. The lack of association of third graders' ratings with the *t*-test values of the data sets may be due to their consistently high confidence ratings. However, for the older participants, it indicates that even when participants were not systematically using the individual characteristics, they were able to use them together to draw conclusions, regardless of their level of explicit access to this information. The fact that the college students showed a small trend in being more confident with increased within-group variability while

demonstrating strong evidence of drawing conclusions closely aligned with *t*-test results may indicate that for these data sets, the overall within-group variability was relatively small (even with our high-variability manipulation). Therefore, reasoning that focused on sample size and between-group variability may have been a reasonable way to draw conclusions about these data.

The participants not only used data characteristics in drawing conclusions but also justified their conclusions explicitly. In doing so, nearly every participant mentioned at least one data characteristic, though the college students referred to a wider range of characteristics and referred to these characteristics more frequently. The third graders appeared to be much less explicitly aware of concepts such as sample size and within-group variability, and the majority of their data-related comments were about trends in the data. Nonetheless, about half of the third graders referred to between-group variability and about half referred to the magnitude of differences between groups. The mention of these characteristics at all at such a young age indicates an early awareness that these features are important ones to consider in drawing conclusions from data.

In addition to the data features, even when we set up a situation with no clear mechanistic explanation, about half of the participants at each age brought their background knowledge to bear and mentioned potential mechanisms for the results that were not explicitly mentioned in the cover stories. Therefore, all students did not evaluate the data in a theory-neutral way. For some, predictions appealed to patterns in the previous data sets. For other students, predictions appealed to theoretical reasons for *why* such patterns were found.

One possible limitation of this study is that the presentation of data might have influenced reasoning by either providing too much information at once or guiding interpretation. Format of problem presentation has been shown to affect children's ability to solve math problems (e.g., Klein & Bisanz, 2000). If this amount of information exceeds processing limits, participants might use nonoptimal strategies for analysis. For example, instead of comparing the entire range of data at once, participants might draw a conclusion from the last set of numbers alone, ignoring the remaining data. In addition, for each trial in Experiment 1, participants were presented two data sets and asked to compare them. By presenting the data in this way, it is possible that participants would only compare *pairs* of data points instead of comparing the entire columns of data. To examine these issues, we conducted a second experiment.

Experiment 2

In the follow-up experiment, we explored how style of data presentation might influence children's and adults' conclusions. We made several specific changes to the methodology from Experiment 1. We changed the data presentation style so that the data were presented incrementally instead of simultaneously. This change in presentation allowed us to reduce the information processing burden and to compare the effect of presentation style on participants' processing of the data. Specifically, we wanted to explore whether the change from simultaneous to paired presentation would highlight different data characteristics (e.g., sample size) and obscure others (e.g., between-group variability). In addition, would a presentation format that prevents direct paired comparison lead to differences in reasoning about data characteristics? If understanding of data characteristics is robust, then the presentation format should have a minimal effect on using the characteristics to draw conclusions. However, if participants were only looking at subsets of the data, then we would expect the pairwise condition to lead to more confidence in evaluations derived from smaller samples. If participants were relying on individual pairwise comparisons as part of their reasoning, then the pairwise condition would likely lead to reasoning more clearly based on between-group variability than the column condition, in which the opportunities for such comparisons were limited. We also predicted that the two formats would lead to increased attention to the data characteristics, particularly to sample size, as the feature that most obviously changed between iterations.

We made other minor modifications in response to observations of participants in Experiment 1. We wanted to find out how likely participants were to want the engineers to test the balls again, an assessment of whether participants believed that they needed additional data to draw a conclusion. From the Experiment 1 cover story, one could argue that there is always a reason to be conservative in conclusions and state that there is not really evidence for a difference (a bias toward a Type II error rather than Type I error). Therefore, we modified the cover story to indicate more clearly that there are costs to being wrong in either assuming that balls go the same distance when they do not or assuming that they do not when in fact they do. This change was to reduce the possibility of participants suggesting simply producing and testing more balls without first considering the available data.

In addition, we changed some of the specific questions to be more explicit about exactly what we

were looking for. In Experiment 1, we asked participants what the engineer had learned from testing, and nearly all participants gave answers about which ball went farther. If they did not, the experimenter probed for that answer directly. In Experiment 2, we changed the question to ask explicitly which ball went farther.

Finally, we added a practice problem that included three presentations of data and one reversed data pair to familiarize all participants with the nature of the task. The practice problem served to introduce participants to the notion that they would be seeing more data after the first trial and also that pair reversals in the data sometimes occur.

Method

Participants. Twenty-two third-grade students (mean age = 9.0 years, range = 8.2–10.6 years), 29 sixth-grade students (mean age = 11.9 years, range = 11.1–13.8 years), and 50 undergraduate students (mean age = 19.8 years, range = 17.8–23.6 years) participated in this study. None of these students participated in Experiment 1. The children were students in two elementary schools and two middle schools in the Northeast United States. The third-grade sample consisted of 54% White students, 21% Asian students, 18% Black students, and 9% Hispanic students. The sixth-grade sample consisted of 56% White students and 44% Black students. The third- and sixth-grade students were from different regions in the same metropolitan area than those who participated in Experiment 1. The college student sample consisted of 62% White students and 38% Asian students.

Procedure. Because there were no differences in the main conclusions drawn based on cover story, all participants in Experiment 2 were shown the robot cover story used in Experiment 1, with following paragraph added:

> It's very important that the engineers can find out whether the balls are the same or different. Only if the balls are the same and go the same distance can the factory sell them to make money. If they aren't the same then the engineers need to remake the balls, which is expensive and time-consuming. So the engineers need your help in looking at the results and deciding if they can be sure whether the balls are the same or different.

Participants were presented with three data sets, each with a different ball type, in one of two presentation formats. The three data sets were identical to those used in Experiment 1, and each included six pairs of data. One set had no reversed pair, one had one reversed pair (the third data pair was greater in Column B than in Column A), and one had two reversed pairs (the second and fifth data pairs were greater in Column A than in Column B). The order of presentation was counterbalanced using a Latin Square design, such that one third of the participants in each age group saw each data set first. All the data used in this study were from the high-variability (two-digit) condition from Experiment 1, as we found no clear effect of within-group variability on reasoning.

In the pairwise condition, participants saw data presented in two columns. First, they were shown one pair of data points, then two pairs, then four pairs, and then six pairs. After each presentation of data, participants were asked if there was a difference between the two balls, how sure they were of whether there was a difference and why, and whether they thought that the engineers should test the balls again. Examples of this condition are shown in Table 3.

Table 3
Example of Data Sets in Experiment 2

Examples of presentation formats shown to participants in the pairwise condition

Basketball A	Basketball B
74 ft	56 ft

Basketball A	Basketball B
74 ft	56 ft
83 ft	65 ft

Examples of presentation formats shown to participants in the column condition

Tennis Ball A	Tennis Ball B
49 ft	64 ft
40 ft	
39 ft	
51 ft	
50 ft	
38 ft	

Tennis Ball A	Tennis Ball B
49 ft	64 ft
40 ft	52 ft
39 ft	
51 ft	
50 ft	
38 ft	

In the column condition, participants saw six data points in one column and one in the other column. They were asked the same questions as in the pairwise condition and were then presented with additional data points in the second data column (one, two, four, and six data points at a time). This condition was included to see if reasoning changes when pairwise comparisons are less salient. Examples of this format are shown in Table 3. Questions asked in both formats were identical, with no explicit mention of the format in the questions.

Results

Judgments of whether there was a difference between the data sets. As in Experiment 1, responses—yes–no and sureness level—were combined into a single 7-point ordinal variable. Because there was no direct assessment of whether participants believed that the ball in Column A or Column B went farther, there was no accuracy measure. Participants stated only whether they thought that there was a difference between the two variables.

To examine the effect of sample size alone, we examined the data set in which reversed pairs were not included. Thus, we compared judgments of whether there was a difference between data sets when participants saw one, two, four, or six pairs of data or one column with six data points and a second with one, two, four, or six data points. These data points were all presented in the same scenario, and data points in one column were always larger than those in the other column. We found that sample size influenced participants' assessments of differences between the data sets. Age and condition were between-subjects variables, and sample size was a within-subjects variable. A 3 (age: third grade, sixth grade, or college) $\times$ 2 (condition: pairwise vs. column) $\times$ 4 (sample size: 1, 2, 4, or 6) mixed-model ANOVA was performed. We found no significant effect of age, $F(2, 95) = 1.73$, $p > .1$; no effect of condition, $F(1, 95) = 1.14$, $p > .1$; and a significant effect of sample size, $F(3, 93) = 4.52$, $p = .005$, partial $\eta^2 = .13$. There was also an Age $\times$ Sample Size interaction, $F(6, 186) = 3.87$, $p = .001$, partial $\eta^2 = .011$. Simple effects analysis to understand the nature of the interaction indicate that college students demonstrate a large, significant increase in sureness with increasing sample size in the pairwise condition, $F(3, 21) = 7.40$, $p = .001$, partial $\eta^2 = .51$, and in the column condition, $F(3, 23) = 17.37$, $p < .001$, partial $\eta^2 = .53$, but that third and sixth graders do not (Fs < 1.0). In addition, there was a significant Sample Size $\times$ Condition interaction, such that the paired condition leads to a linear trend of increasing sureness ratings with increased sample size, whereas the column condition does not, $F(3, 93) = 4.97$, $p = .003$, partial $\eta^2 = .14$. See Figure 3 for means for each age and sample size, by condition.

Between-group variability also influenced assessments of differences between data sets. The ratings of whether there was a difference between groups on the sixth trial (when all data had been presented) were compared across stories in a mixed-model ANOVA. Each participant saw one story with no reversed pair, one with one reversed data pair, and one with two reversed data pairs. A 3 (age: third grade, sixth grade, or college) $\times$ 2 (condition: pairwise vs. column) $\times$ 3 (between-group variability: zero, one, or two reversed data pairs) ANOVA was performed. We found no effect of age, $F(2, 95) = 0.74$, $p > .1$; an effect of condition, $F(2, 95) = 5.59$, $p = .020$, partial $\eta^2 = .06$; and an effect of between-group variability, $F(2, 94) = 12.60$, $p < .001$, partial $\eta^2 = .21$. On average, there were higher ratings of a difference in the pairwise condition and higher ratings with less between-group variability (i.e., with fewer reversed pairs). There was also a significant Age $\times$ Between-Group Variability interaction, $F(4, 188) = 3.87$, $p = .005$, partial $\eta^2 = .08$. Examining the Age $\times$ Variability interaction, only the college students reduced their sureness ratings considerably when there was more between-group variation, in both the pairwise condition, $F(2, 22) = 18.94$,

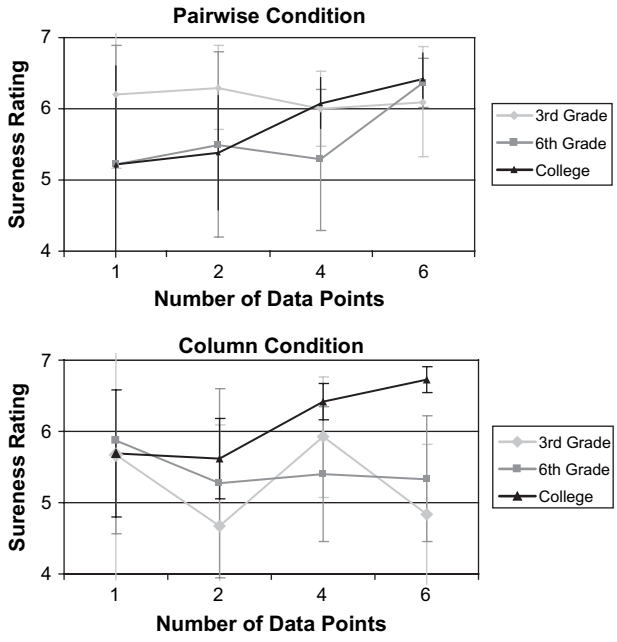

Figure 3. Experiment 2: Participants' sureness ratings at each grade with one, two, four, and six pairs of data but no reversed data pairs (by presentation format).
Note. Error bars represent standard error of measurement.

$p < .001$, partial $\eta^2 = .63$, and the column condition, $F(2, 24) = 20.61$, $p < .001$, partial $\eta^2 = .63$. There was a trend toward an interaction between age and condition, such that college students were more sure in the column condition, but third and sixth graders were more sure in the pairwise condition, $F(2, 95) = 2.37$, $p = .099$, partial $\eta^2 = .05$. The mean sureness ratings are presented in Figure 4 and indicate a clear pattern of decreased sureness ratings with increased variability in college students across both conditions.

As in Experiment 1, we calculated *t* tests for each set of data participants saw. In the pairwise condition, the *t* values ranged from 0 to 3.98. For this condition, we again found evidence of a strong association in college students between average sureness ratings and *t* statistics ($r = .924$, $p < .001$). Third graders exhibited a similar pattern ($r = .685$, $p = .042$). Sixth graders did not demonstrate a clear association ($r = .484$, $p = .177$). In the column condition, a different pattern emerged. These *t* values also ranged from 0 to 3.98, but only the college students showed any evidence of using this information systematically. College students' confidence in the difference between groups was very closely matched to actual *t* statistics ($r = .906$, $p = .001$). In contrast, third and sixth graders' ratings showed no evidence of a positive association between confidence and *t* statistics (third grade: $r = -.085$, $p > .5$; sixth grade: $r = .035$, $p > .5$).

Reasons offered. To classify explanations for reasoning, we began with the same codes as in Experiment 1. Two coders independently coded each set of explanations. The overall agreement level was 86.3%, and all discrepancies were resolved through discussion.

The pattern of explanations offered was similar to that in Experiment 1, with large grade effects. However, there were some key differences between the patterns of responses offered in Experiments 1 and 2. The frequencies of participants mentioning each explanation are summarized in Tables 4 and 5. The incremental data presentation led to more frequent mentions of data characteristics such as sample size (e.g., "The engineer should test again to get more data"; "Don't need to test again because we already have a lot of data"), suggesting that this format increased the salience of this characteristic. In Experiment 1, only 10% of third graders and 27% of sixth graders mentioned sample size. However, when directly comparing data sets with increasing data, 40% of third graders and 71% of sixth graders mentioned sample size in the pairwise condition, whereas 33% and 47%, respectively, did so in the column condition. Nearly all college students mentioned sample size. In discussing variability, mention of reversals or lack thereof was more common in the pairwise condition across all age groups. Mention of spread was similarly uncommon for third and sixth graders in the column condition as in Experiment 1 (8% and 13% as compared to 3% and 15%, respectively), and not a single third or sixth grader referred to spread in the pairwise condition.

Participants were asked to reason about all the data available in each scenario. However, as noted earlier, we found some participants explicitly stating that they only were referring to a single pair of data points in drawing conclusions. In Experiment 1, only 23% of third graders and 1 sixth grader ever clearly referred to a subset of the data in justifying answers. In

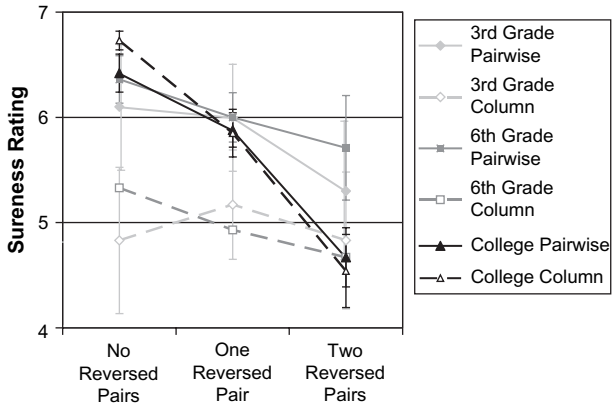

Figure 4. Experiment 2: Participants' sureness ratings at each grade with six pairs of data, by level of between-group variability in the data and presentation format.
Note. Error bars represent standard error of measurement.

Table 4
Experiment 2: Percentage of Participants at Each Age Who Mentioned Each Justification for Their Sureness, in the Pairwise Condition

	Third grade	Sixth grade	College
Data responses			
Sample size**	40	71	100
Reversed pairs**	30	57	83
No reversed pairs*	30	64	83
Spread of data	0	0	21
Overall trend in data	100	93	100
Single pair "trend"**	80	43	13
Magnitude of differences**	20	64	88
Magnitude of single pairs*	30	57	17
Inference**	30	43	88
Mechanism responses			
Ball property	20	36	21
Robot/athlete property*	40	57	13
Environment	0	14	8

Note. Age differences: *$p < .05$. **$p < .01$.

Table 5
Experiment 2: Percentage of Participants at Each Age Who Mentioned Each Justification for Their Sureness, in the Column Condition

	Third grade	Sixth grade	College
Data responses			
Sample size**	33	47	92
Reversed pairs	17	20	46
No reversed pairs**	8	20	69
Spread of data	8	13	35
Overall trend in data	83	87	100
Single pair ''trend''	58	53	31
Magnitude of differences**	17	27	85
Magnitude of single pairs	58	47	38
Inference*	8	33	58
Unequal amount of data	50	53	73
Mechanism responses			
Ball property	8	27	12
Robot/athlete property	25	20	19
Environment	0	7	15

Note. Age differences: *$p < .05$. **$p < .01$.

contrast, the incremental presentation formats in Experiment 2 led many more participants to reason this way. As shown in Tables 4 and 5, many participants in all age groups referred to a pattern or trend in a single pair of data points or to the magnitude of the difference between the two numbers of a single pair in justifying their conclusions (e.g., after seeing the final two numbers presented in the column condition, referring to only these two numbers: ''52 is far away from 40''; similar patterns occurred after seeing the last pair in the pairwise condition). In fact, the only significant age difference in Experiment 2 on these issues was in the pairwise condition, where 80% of third graders, 43% of sixth graders, and 13% of college students referred to a ''pattern'' or ''trend'' based on a single pair of data to justify their responses, $\chi^2(2, N = 48) = 14.54$, $p < .001$.

We also expanded the codes used in Experiment 1 to include more codes for data-based responses. In the column condition, it was common for participants to justify their request for more data with a reference to the fact that the columns were unequal; thus, it was common for participants to say that they would like more data in Column B because there was currently not as much data as in Column A (e.g., ''There's four more numbers on A and there still needs to be four more on B so it will be even,'' ''You need even amounts of data''). Therefore, we noted whenever participants referred to unequal sample sizes. In addition, because we asked participants to assess whether they wanted more data, many participants used the data to make inferences about whether more data were needed. We added the code of inference to account for answers such as ''It's pretty clear that they're different'' and ''More data will confirm it one way or another.''

In addition to discussions of data, we also explored whether participants offered causal mechanistic explanations as justifications for their conclusions. As in Experiment 1, we coded explanations for a mention of characteristics of the robot, characteristics of the ball, and characteristics of the environment. Overall, a little more than one third of the participants mentioned at least one mechanistic explanation (e.g., ''Because while the robot threw the first two balls, he may have lost some power''; ''There might be damage or air leakage in the ball''), with no significant age differences (36% of third graders, 45% of sixth graders, and 34% of college students mentioned such an explanation).

Testing again. In Experiment 2, participants were asked if they thought the engineers should test the balls again. If they said yes, they were asked whether the engineers should test Ball A, Ball B, or both. With only one data pair, participants asked for more data 93.4% of the time; with two data pairs, participants asked for more data 90.1% of the time; with four data pairs, participants asked for more data 71.6% of the time; and with six data pairs, participants asked for data 40.9% of the time. The grade breakdown showed an interesting pattern after all six data pairs were presented. Third graders wanted to test for more data about 60% of the time across between-group variability levels (zero, one, or two overlaps). Sixth graders wanted to test for more data about 30% of the time across levels. College students, however, were most likely to differentiate among the data sets. Thus, although only 14% of the college students asked for more data when there was little between-group variability, 40% asked for more data when there was one overlapping pair and 56% asked for more data when there were two overlapping pairs. See Figure 5 for details.

Discussion

Experiment 2 was conducted to determine if the relative salience of various data characteristics could be manipulated by varying presentation format. The results suggest that presentation did not change the pattern of explicit reasoning shown in Experiment 1 for college students. The majority of college students clearly changed their ratings systematically with sample size. There were no differences in ratings of sample size based on whether the data were presented pairwise or in columns, indicating that the processing of each format was comparable for sample

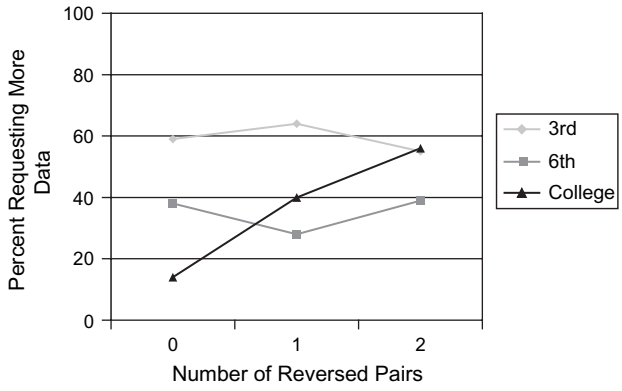

Figure 5. Experiment 2: Percentage of participants who requested further testing after six trials, based on between-group variability. *Note.* The data are broken down by age.

size ratings. However, third and sixth graders did not systematically use sample size when viewing data in these formats. This finding suggests that presenting data in this incremental format may in fact have made it somewhat more difficult for children to systematically use sample size when data sets they reasoned about were being updated.

However, the patterns of explicitly discussing sample size differed from Experiment 1. In Experiment 2, far more participants in the third and sixth grades talked about sample size as affecting their judgment. The presentation format increased the salience of this characteristic enough for many children to notice that sample size might play an important role. At the same time, incremental presentation of data led to more frequent references to trends in single pairs of data or subsets of the data instead of the whole pattern, though this trend was more pronounced for the younger participants. Because the task involved adding only one or two data points with each new trial, it is not surprising that this feature encouraged more participants to focus solely on a subset of the data rather than the whole, particularly in the pairwise condition.

Assessments of variability were also affected by data presentation format. College students demonstrated clear patterns of decreased sureness when there was more variability in the data, whereas the responses of younger participants were not as systematic. In addition, the format changes led to some differences in the explicit discussion of variability-related issues. In particular, participants appeared to be more likely to mention a lack of reversed pairs in the paired condition than in the column condition or in Experiment 1. Mention of spread of the data did not vary significantly by age, but mention of reversed pairs (or lack thereof) generally did. These findings suggest that children have a nascent understanding of variability but one that is not yet fully formed and applied equally across contexts.

General Discussion

The results of our study demonstrate that (a) children and adults attended to sample size and between-group variability when drawing conclusions from data; (b) there were significant age-related increases in the detection, use, and explicit awareness of individual data characteristics; and (c) the data presentation format affected the use of these characteristics. This research is the first to specifically examine the use of data characteristics outside knowledge-rich contexts and the ways these characteristics affect reasoning about data.

The framing in Experiment 1 allowed us to present two conditions (athlete, robot) that provided two different sources of error variance (e.g., human variation and robot variation, with the expectation that robots were likely to be perceived as less variable than humans). Although there were more mechanism-based explanations in the athlete condition than in the robot condition, there were few differences in how the data characteristics of sample size and variability were interpreted, suggesting that these characteristics were a source of information separate from domain knowledge. The fact that nearly every participant explicitly mentioned at least one data characteristic in justifying conclusions indicates an awareness of the importance of the numerical values. At the same time, Experiment 2 demonstrated that incremental data presentation led to more explicit reasoning about different data characteristics. These results suggest that participants may have difficulty reassessing sets of data with incremental additions to the data, even as these incremental additions make awareness of sample size more explicit. This pattern may indicate that initial values are disproportionately weighted relative to later values, though additional research is needed to evaluate this possibility. The most salient difference between presentation formats in Experiment 2 was in reasoning about between-group variability. In the column condition, participants still tended to use the characteristic but were slightly less sure of their conclusions, suggesting that the break from a focus on pairs often made them reason about the data set differently. However, when participants failed to consider the whole data set, they most often appear to have looked only at the most recently presented data and commented on a direct comparison on only this subset of the data.

The results demonstrate that children's data evaluations can be influenced by the characteristics of the data when comparing sets of data. But through what processes are reasoners making these comparisons? Though we did not specifically test this model, one possibility is that a mechanism for magnitude comparison and an explicit evaluation process may work in concert to produce the data interpretation described previously. Case and Okamoto (1996) provided evidence for a developmental model in which implicit processing mechanisms provide the foundation for children's learning in various domains such as number (i.e., central conceptual structures). Initially, development is highly constrained by processing limits that in turn constrain the number of dimensions a child can process, but as space and knowledge increase, children can attend to (and simultaneously process) multiple dimensions. Knowledge becomes increasingly available for explicit processing and can be extended to other areas (e.g., knowledge of living things). This model is one of many that suggest implicit processing mechanisms working in concert with gradually developing, explicit processes to account for developmental change within complex domains (e.g., Feigenson et al., 2004).

Our participants may have compared sets using approximate means and variances derived from the data. Though there is no definitive account of numerical estimation, one well-documented explanation suggests that number is represented by analog magnitudes that contain a proportion of error (as opposed to exact quantities; Gallistel & Gelman, 2000). There is evidence that estimating ability develops slowly and is difficult for many children and even adults (for discussion, see Siegler & Booth, 2005). Further evidence from recent research with adults suggests that estimating magnitude may be a general-purpose mechanism, showing similar patterns across varied information sources (Barth, Kanwisher, & Spelke, 2003). Although these studies presented data sets in a one-time comparison (not a series of data points to be combined later), our results suggest that the same representational mechanism used to represent and compare single quantities may represent and compare mean values. Obrecht, Chapman, and Gelman (2007) provided adults with a series of product ratings and found that participants were more confident in differences as mean differences increased and variance decreased, as in a *t* test.

If the process of mean computation is relatively automatic, then differentiating individual dimensions may require the acquisition of strategies that allow their identification and representation. Processing capacity (e.g., space, speed) likely plays a large role in the age-related performance differences (Gathercole et al., 2004; Kail, 2007). The transformation from presentation format (e.g., a set of values) into a different representational format (e.g., rough mean with error variance) requires sufficient space to implement these operations (Halford et al., 2007). In fact, this may be why children attended to sample size with simultaneous presentation (as in Experiment 1) but not with incremental presentation (as in Experiment 2). Research on processing capacity (e.g., Gathercole et al., 2004) indicates significant differences in working memory capacity between third and sixth graders and may help explain the age-related differences in attending to multiple data dimensions.

In addition, domain knowledge is likely necessary to identify individual data characteristics. Domain knowledge provides information about which dimensions may be associated with differences in behavior or performance. For example, if a professional golfer drives only 100 yards, most golf fans would ascribe this outcome to factors other than ability (e.g., wind, losing balance, losing concentration), whereas the same drive for a novice golfer might be ascribed to ability. In the present experiments, the domain knowledge was limited and was therefore only explicitly used some of the time in justifying conclusions. Setting the same data in a context in which participants have very strong prior theories may have led to different results and much greater reliance on this background knowledge (e.g., if we told participants Golfer A was a professional golfer and Golfer B was a complete beginner).

Related to domain knowledge is knowledge about using data themselves, particularly in understanding the relationship between a sample and a population. Increasing sample size increases the approximation to the population under study. Younger children in the present experiments did not appear to treat data as samples, as there was little effect of sample size. Paradoxically, third graders were often more certain with a single observation than with multiple observations. Perhaps, a source of young children's difficulty is an assumption that one observation is indicative of a population.

The results of this study suggest a prominent role for data in children's reasoning. Absent strong background knowledge, children make use of data as a source of information. Future research should more closely examine the influence of processing factors such as capacity and speed, in addition to varied representations. Future research should also examine how children's use of data (and data characteristics) might be useful for acquiring knowledge in novel domains.

Conclusions

In sum, these findings demonstrate that children and adults attend to the number and variability of observations when interpreting data. The results also demonstrate that the use of these characteristics improves with age and becomes more explicitly accessible. Response patterns are consistent with an intuitive *t* test, in which reasoners represent and compare an approximate mean that includes information on the variability of observations. With experience, children better identify the individual dimensions (e.g., variability), attend to finer distinctions of these dimensions, and increase their explicit access to this information. Although children's background knowledge is an important factor in interpreting data, our results indicate that the characteristics of the data themselves play a role in how children make sense of observations.

References

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, *3*, 201–221.

Case, R., & Okamoto, Y. (1996). The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development*, *61*(1–2).

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.

Echevarria, M. (2003). Anomalies as a catalyst for middle school students' knowledge construction and scientific reasoning during science inquiry. *Journal of Educational Psychology, 95*, 357–374.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307–314.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences*, *2*, 59–65.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology, 40*, 177–190.

Gutheil, G., & Gelman, S. (1997). Children's use of sample size and diversity information within basic-level categories. *Journal of Experimental Child Psychology, 64*, 159–174.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences*, *11*, 236–242.

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

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

Kail, R. V. (2007). Longitudinal evidence that increases in processing speed and working memory enhance children's reasoning. *Psychological Science*, *18*, 312–313.

Klaczynski, P. A., & Aneja, A. (2002). Development of quantitative reasoning and gender biases. *Developmental Psychology, 38*, 208–221.

Klein, J. S., & Bisanz, J. (2000). Preschoolers doing arithmetic: The concepts are willing but the working memory is weak. *Canadian Journal of Experimental Psychology, 54*, 105–116.

Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology, 64*, 141–152.

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

Koslowski, B., Okagaki, L., Lorenz, C., & Umbach, D. (1989). When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development*, *60*, 1316–1327.

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

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, Serial No. 245).

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

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

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction*, *22*, 219–290.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive t-tests: Lay use of statistical information. *Psychonomic Bulletin and Review, 14*, 1147–1152.

Piaget, J., & Inhelder, B. (1975/1951). *The origin of the idea of chance in children*. New York: W.W. Norton.

Rubinsten, O., Henik, A., Berger, A., & Shahar-Shalev, S. (2002). The development of internal representations of magnitude and their association with Arabic numerals. *Journal of Experimental Child Psychology, 81*, 74–92.

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

Siegler, R. S., & Booth, J. L. (2005). Development of numerical estimation: A review. In J. I. D Campbell (Ed.), *Handbook of mathematical cognition*. New York: Psychology Press.

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

Temple, E., & Posner, M. (1998). Brain mechanisms of quantity are similar in 5-year-old children and adults. *Proceedings of the National Academy of Sciences USA*, *95*, 7836–7841.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*, 99–149.

Zimmerman, C. (2007). The development of scientific thinking in elementary and middle school. *Developmental Review*, *27*, 172–223.