Making Sense of Data: Identifying Children’s Strategies for Data Comparisons

Bradley J. Morris, Amy M. Masnick, and Christopher A. Was

Kent State University, Kent, OH, USA; Hofstra University, Hempstead, NY, USA

ABSTRACT
The statistical properties of data are not present in any individual value, but rather, emerge only by perceiving the set as a whole. Summarizing the statistical properties of sets (e.g., creating ensembles) is ubiquitous in cognition, yet one unanswered question is how this process changes over development. The properties of number sets (e.g., means) provide a unique opportunity to investigate the mechanisms underlying summarization. We presented fourth (~ten-year-old) and sixth grade (~twelve year-old) children from the Midwestern region of the United States with a data comparison task, determining which of two golfers produced the farthest drive, and measured their accuracy, confidence, and eye fixation patterns while solving each trial. Children’s data strategies were identified by coding their eye tracking patterns. The results demonstrated that accuracy and confidence were related to the statistical properties of the sets. Older US children consistently used a strategy that demonstrated attention to diagnostic set properties (e.g., attending to most numbers in a set), whereas most younger children used a variety of strategies, many of which were less accurate (e.g., attending to only one number in a set) or used the same strategies less efficiently than older children (e.g., attending to non-diagnostic place values). The results add to our understanding of US children’s quantitative reasoning by identifying strategies children use to make sense of data, their developmental transitions, and how changes in children’s strategy use is a key component in understanding the developmental improvements in summarizing complex information in the environment.

How do children make sense of data, or numbers in context? Understanding data and statistics is important in daily life and in math education, as outlined in the Common Core standards (National Governors Association, 2010). Making sense of data involves attending to the properties of individual numbers in a set (e.g., magnitude) and the properties of the set not present in any single number (e.g., means). Detecting set properties requires a representation of the set itself, likely through summarization (Masnick & Morris, 2022). There is some evidence that children can summarize the statistical properties from information in their environments (e.g., Rhodes et al., 2016) and summarization is an adaptive way for children to make sense of the environment while reducing the burden on their limited processing capacities (Whitney & Yamanashi...
Leib, 2018). However, open questions include documenting what strategy or strategies young children use to summarize these properties from data sets, if children use effective strategies with low fidelity, and if children initially use a variety of strategies to make sense of data and show a developmental improvement in the selection and implementation of different strategies. The goal of the present experiment was to investigate how children make sense of data by identifying their strategies using eye tracking patterns and evaluating the efficacy of these strategies and their implementation as a means of understanding the development of children’s reasoning about data evaluation.

**Summarization**

Adults frequently summarize the statistical properties of complex information in their environments (Heydt et al., 2012; Wagemans et al., 2012; Whitney & Yamanashi Leib, 2018). For example, when shown a series of dots, adults summarize the average size of the set of dots, as demonstrated by erroneously “recognizing” the average more frequently than actual dots from the set (Ariely, 2001). Although there is considerable evidence that adults summarize such information rapidly (Obrecht, 2019; Peterson & Beach, 1968), there is surprisingly little evidence regarding children’s ability to summarize statistical properties. Children and adults can quickly apprehend the statistical properties of sets of faces (e.g., average happiness of a set of faces; Rhodes et al., 2016) and children as young as 4 can quickly summarize the average size of an object from a set of objects, though not as rapidly as adults (Sweeny, Wurnitsch, Gopnik, & Whitney, 2015). When 4- and 5-year-olds were shown groups of oranges in trees, they were able to quickly summarize the average size of each set of oranges, though adults were more accurate (~75% vs. ~95%, respectively (Exp. 1); Sweeny et al., 2015).

It is possible that children summarize the statistical properties of number sets, much as adults summarize the statistical properties of complex visual scenes (Wagemans et al., 2012; Whitney & Yamanashi Leib, 2018). There is evidence that children (like adults) can make rapid and accurate informal estimates from data sets (Masnick & Morris, 2008; Watson & Moritz, 1999, 1998). For example, when shown two sets of multi-digit numbers, children as young as nine quickly estimated which set of data was larger than the other at above chance levels, though at significantly lower levels than adults (Masnick & Morris). However, little is known about why summaries improve over development and how this develops within the domain of number.

**Making sense of data**

Making sense of numerical data begins with making sense of individual numbers. Individual numbers are represented as precise, verbal categories (e.g., “15”) and as approximate, relative magnitudes with error variance (Dehaene, 2009). Approximate representations are activation functions on a mental number line that peak near a value and decline with increasing distance from the value (Park & Starns, 2015). Comparisons between single-digit numbers are faster and more accurate as the ratio of difference between numbers increases (i.e., distance effect; Buckley & Gillman, 1974; Moyer & Landauer, 1967; Sprute &
Temple, 2011). Approximate representations appear to be used for these comparisons because accuracy is greater and solutions are faster as the spatial distance between approximate representations on a mental number line increases (Nieder & Dehaene, 2009).

Number sets have unique properties, not present in single values such as central tendency, variance, and set size. Obviously, these properties can be formally calculated, but they can also be estimated informally. In adults, comparison accuracy improves as the difference between the set means (e.g., the ratio of means) increases (Irwin & Smith, 1957; Morris & Masnick, 2015; Obrecht, Chapman, & Gelman, 2007; Peterson & Beach, 1967; Trumpower, 2013). For example, when comparing products on the basis of aggregated consumer ratings, confidence judgments were most affected by mean differences in ratings (Obrecht et al. 2007). When estimating the number of dots in a subset of two mixed sets of dots, participant responses were influenced by the ratio of set means (Cordes, et al., 2013). When asked which golfer hit the ball farther, adults were more accurate and more confident given a high (4:5) than a low (9:10) mean difference, and produced more visual fixations when comparing sets with low mean differences than high mean differences, suggesting that smaller differences cue the search for more information (Morris & Masnick, 2008). Variance is also perceived intuitively (Lathrop, 1967); however, it is often not as salient as mean differences (Masnick & Morris, 2008; Obrecht et al.; Trumpower, 2013). The interaction between these properties contribute to the distinctiveness of sets. Specifically, mean difference and variance interact such that comparisons and confidence were more accurate between sets with high mean differences and low variance than those with low mean differences and high variance (Morris & Masnick, 2015; Obrecht et al., 2007; Peterson & Beach, 1967; Trumpower, 2013).

One explanation for rapid, accurate performance is that numbers in a set have both individual representations and are part of summary representations that include means and variance. Comparisons can be made on the basis of the difference between summary representations. Summary representations could emerge from multiple approximate number representations (Masnick & Morris, 2022; Morris & Masnick, 2015). Recall that individual numbers are associated with an activation function on a mental number line that peaks around a specific location and declines proportionally with distance from this location (Bonny & Lourenco, 2013). Multiple numbers would create multiple activations, resulting in summaries that represent approximate set means and variance. Activation areas could differ in the amount of within-set similarity (approximate mean) and the width of the activation area (approximate variance; see Alvarez, 2011; Morris & Masnick, 2015 for a more detailed explanation). During comparisons, summary mean activations should be more salient than variance. A highly dissimilar set of numbers would result in an indistinct area of activation and a wide overall activation area while a highly similar set of numbers would produce a distinct area of activation and a narrow overall activation area. The summary activation areas would be easier to distinguish given distinct differences (i.e., distances) between them.

As stated above, statistical properties that are not present in any individual set member emerge from a set of numbers. When perceiving sets of numbers or objects, the whole is greater than the sum of its parts. We refer to this process as summarizing statistical properties, while acknowledging and incorporating the findings from similar approaches (e.g., deriving Gestalt impressions, Wagemans et al., 2012, and creating ensemble codes, Whitney & Yamanashi Leib, 2018). One must attend to a sufficient proportion of the
information in a set to summarize the statistical properties with reasonable fidelity (Alvarez, 2011). Although summaries of these statistical properties are rapid and largely accurate, the direction of attention influences the nature of and quality of the representation (Whitney & Yamanashi Leib, 2018). For example, focusing attention on specific set members (e.g., a small dot), influences the subsequent representation (e.g., underestimating the set average; De Fockert & Marchant, 2008). This pattern suggests that attention to set elements determines what will be included in the summary value. For example, only attending to one number in each set will provide single number comparisons. However, attending to the hundreds place value of all numbers in two sets will yield two summary values that can then be compared. Although there is considerable evidence that adults and children summarize the properties of object sets (e.g., dots), less is known about the underlying processes of summarizing numerical data.

There are many ways to make sense of data; however, these strategies for data comparisons differ in their utility. For example, during comparisons, when data sets are very different from each other (e.g., very large mean difference), nearly any strategy should yield a correct answer. However, when data sets are similar, children will need to pick up a sufficient amount of diagnostic information to detect a true difference between sets. We suggest that there is an expected strategy that is most effective in picking up information when solving this task, and we discuss what makes a strategy expected in detail below. We will investigate these underlying processes by identifying the strategies used by children to make sense of data and evaluate the degree to which strategies are used effectively.

The role of strategies in summarization

The goal of the current experiment is to identify the data strategies children use from real-time, behavioral data created through process tracing, that is, from the eye tracking patterns during problem solving (Schulte-Mecklenbeck et al., 2017). Strategies have been investigated in a variety of domains such as spelling (Rittle-Johnson & Siegler, 1999), science (Chen & Klahr, 2008), deductive reasoning (Morris & Sloutsky, 2001), mathematics (e.g., addition; Thevenot, Barrouillet, Castel, & Uittenhove, 2016), and playing games (Crowley & Siegler, 1993). Children adaptively select and modify their strategies with experience (Siegler, 2016) and learn to discard less effective strategies over time (Kuhn, 2013). This process is dynamic and changes over very small time scales, as demonstrated by a child using multiple strategies across trials within the same task or within a single trial (Siegler, 2016).

The overlapping waves model provides a conceptual framework to explain how strategies are acquired, their outcomes evaluated, how new selections are made, and why children demonstrate within- and between-task variability (Siegler, 2005, 2016). An earlier model, the test-operate, test-exit model, demonstrated that a strategy is implemented, feedback on its efficacy is obtained, and if the goal is not attained, an additional attempt is made using either the same or a different strategy (Miller, Galanter, & Pribram, 1960). In the overlapping waves model, strategy selections are adaptive in that children tend to select the fastest and most accurate strategy available (Fazio, DeWolf, & Siegler, 2016; Siegler, 2016). When a strategy does not produce an expected result, this prompts additional steps in the form of using the same strategy again or a different strategy to achieve the processing goal, even in the absence of feedback (Alibali, 1999). There is evidence that children demonstrate
multiple, conflicting ideas about and approaches to making sense of data, focusing on individual elements for some problems and relational elements for others (Watson & Moritz, 1999). Further, children and adults may select a strategy but use it ineffectively, resulting in incorrect answers (Miller, 2000). The overlapping waves model may help explain why children demonstrate such variability in their strategy decisions in data interpretation.

It is also possible that children do not initially use expected strategies to summarize the statistical properties of number sets but acquire these strategies over a developmental and educational progression. As a result, children may initially use a variety of strategies and increase their use of expected strategies over the course of development (Siegler, 2005, 2016). When asked to summarize an average (i.e., determine the center of a series of dots), children used different strategies than adults to solve this task and these strategies were often less effective for producing a correct response (Jones & Dekker, 2018). In this case, children used a variety of strategies within the same session to estimate the center of a cloud of dots. Thus, at least in some domains, children used less effective strategies for summarizing the statistical properties of sets. The use of less effective strategies may be driving less accurate summary estimates.

Another possible factor limiting performance is the fidelity of strategy use. Children’s ability to differentiate perceptual and conceptual information improves as they attend to and summarize diagnostic information in the environment (Gibson, 1969). More recent work from this tradition suggests that performance in high-level cognitive tasks such as mathematics is related to learning to focus attention on diagnostic patterns (Goldstone et al., 2010). For example, when shown algebra equations, children initially focus on obvious features such as numbers versus letters but later attend to more nuanced relations such as the order of operations (e.g., items in parentheses are attended to first; Goldstone et al., 2017).

The limited evidence related to detecting diagnostic patterns in data sensemaking suggests that children may initially focus on irrelevant information (Garfield & Ben-Zvi, 2007; Hancock, Kaput, & Goldsmith, 1992). For example, third and fourth grade children may initially focus on individual values as being representative of the set (Watson & Moritz, 1999), leading them to incorrect answers. Older children in this study were more likely to focus on the entire set when drawing inferences from data (see Hancock et al., 1992 for similar results).

**Identifying data sensemaking strategies**

Our goal was to identify children’s strategies for comparing number sets to elucidate the underlying processes used to summarize the statistical properties of number sets. In particular, to better understand how children make sense of data, we aimed to identify children’s strategies, evaluate the efficacy of strategy use, measure the variability between and within participants, and assess the accuracy of different strategies. We used eye tracking because it provides real-time, direct evidence to what information is being attended and at what point in the reasoning process. Eye tracking has been described as process tracing (Hayhoe & Ballard, 2005) because visual attention is gathered slightly in advance of action and attention allows organisms to monitor their actions. For example, in a study of a participant making a sandwich, nearly all fixations preceded actions (e.g., looking at jar
of peanut butter before reaching) or occurred during actions (i.e., monitoring; Rothkopf & Pelz, 2004). Mapping the pattern of what children are attending to during the task allows us to identify strategies children used when comparing data sets. We identified possible strategies based on pilot data and from coding data from previous experiments (Morris & Masnick, 2015). As discussed above, an expected strategy would include attention to a sufficient sample of the information in a set (e.g., a large enough sample to approximate the properties of the set itself) as well as attention to diagnostic features of sets (e.g., hundreds values in three-digit numbers). The use of other strategies may provide information about how naïve children (i.e., children who have received no formal instruction) make sense of data. Table 1 provides example data sets, framed as the results of a golfing competition in which two golfers (Golfer A and Golfer B) are competing to win a contest for the longest drive (i.e., a golf shot with the goal of producing maximum distance). In this example, each golfer hits a ball five times and the result is measured in feet. The goal was to determine which golfer hit the ball farther and a correct response was operationalized as choosing the set with the higher mean. Based on previous research (Masnick & Morris, 2008; Morris & Masnick, 2015), we hypothesized specific strategies and their associated eye tracking response patterns below (see Table 2).

| Table 1. Example data sets. |
|-----------------------------|
|                             |
| **Example A**               |
| Golfer 1 | Golfer 2 | 220 | 170 |
| Golfer 3 | 155 | 430 |
| 150      | 200 | 175 |
| 175      | 210 | 180 |
| 150      | 160 | 190 |
| 178      | 170 | 130 |
| **Example B**               |
| 174.6 | 182 | 166 | 451 |

*Note. Participants do not see the mean.*

| Table 2. Proposed strategies and descriptions of their associated eye tracking patterns. |
|----------------------------------------------------------------------------------------|
| **Strategy** | **Comparison** | **High-Low** | **Gist** |
|-------------|----------------|--------------|----------|
| Description | Comparisons are made from scanning a subset of the total set (e.g., first two, last two). | Comparisons are made between the highest (or lowest) values in each set. | Comparisons are based on estimates of set properties (e.g., mean, variance). |
| Fixation Pattern | Fixations on a limited part of the available set (e.g., first two values only). No more than two per set. | Search for high or low value; Scan set of whole numbers that ends on high/low value; possible iterative fixations between these values. | Sequential scan of all numbers or majority of numbers in each set. Scan can be limited to specific place values (e.g., hundreds). |
| Relative Duration | Overall time per comparison is brief (typically >1500 MS) | Overall time per comparison moderate (typically >1500 MS but can be > 5000 MS) | Overall time per comparison is brief (typically < 1500 MS) |
| Visual Example | ![Visual Example](image) | ![Visual Example](image) | ![Visual Example](image) |

*Note. We also coded for a calculation strategy in which a participant explicitly calculates values such as sums or means. Only one child used this strategy and we eliminated the strategy from further analysis. More information about this strategy is included in the Supplementary materials.*
Strategy 1: Comparison. One possible strategy is to compare a limited part of the total set of numbers (e.g., comparing only the first two drives in a set), prompting comparisons between potentially unrepresentative samples. In the example in Table 1, if a student compared only the first two numbers, she would erroneously choose Golfer 1 in Example A and correctly choose Golfer 4 in Example B. Focusing only on an unrepresentative sample of values, rather than the entire set, ignores critical information about the set. Use of this strategy may suggest a belief that any value equally represents the properties of the entire set. The evidence presented above (Masnick & Morris, 2008) suggested that this is a common approach for younger children (e.g., elementary school).

Strategy 2: High-Low. Another possible strategy is to compare the highest or lowest value in each set, which requires searching through each column for the highest (or lowest) value, and then comparing these two values. Returning to the examples in Table 1, if our student used this strategy, she would incorrectly choose Golfer 1 for Example A and correctly choose Golfer 3 in example B. The use of this strategy may indicate knowledge about sets in which certain values, e.g., the highest or lowest values, are more diagnostic than other values. Although this strategy is more sophisticated than the comparison strategy above, it potentially ignores relevant information about the other properties of the set (e.g., differences in means, variance).

Strategy 3: Gist. Another strategy for comparing sets of numbers is to scan the entire set for an approximation of set properties, such as a rough estimation of mean and spread (see Morris & Masnick, 2015 for a discussion). Approximate averaging has been demonstrated in perceptual research; for example, presented with sets of dots of different sizes, adult participants appeared to encode the properties of the set of dots. When asked to select a dot they had seen, and shown a circle from the set and an “average” circle, that is, a circle that was a mathematical average of the size of the dots in the set and not in the original set, participants chose the average more frequently than they selected actual members of the original set (Ariely, 2001; Chong & Treisman, 2003, 2005). In these studies using dots, the gist of the set was an average of the size of the dot array. As noted above, there is evidence that the gist of a set of numbers may similarly yield an approximate, mathematical average (Morris & Masnick, 2015). It was assumed that participants generated approximate averages (i.e., the gist) of the set, not via the deliberate strategy of averaging, but instead from the intuitive (c.f., Dehaene, 2009) representation of a set of values based on the very fast reaction times (often less than 100 MS). The gist of a set, or an approximate summary of properties such as the mean of a set (Brainerd & Reyna, 2004), can be obtained by scanning entire numbers. For example, for a set of three-digit numbers, a student who encodes only the values in the hundreds column will encode a more diagnostic gist than a student who encodes only the ones column (Morris & Masnick, 2015). Children decompose multi-digit numbers early in development, as demonstrated by comparing compatible number pairs (i.e., when both digits of one number are bigger than in the comparison number; 68 vs. 25) faster than incompatible number pairs (e.g., 51 vs. 37, where 5 is bigger than 3, but 1 is smaller than 7; Nuerk, Weger, & Willmes, 2001). However, a child who scans only a small proportion of numbers or focuses on non-diagnostic place values would likely create an unrepresentative summary. Table 2 provides detailed descriptions of the eye tracking patterns used for coding.
Low fidelity strategy use. Another factor is that strategies may be used but used with low fidelity, i.e., ineffectively. One possible explanation is that younger children were not using strategies effectively (Miller, 2000). There is evidence that children (Bjorklund et al., 1992) and adults (Gaultney et al., 2005) sometimes implement strategies with low fidelity. However, it is also possible that a child could be using a partial strategy (Waters, 2000). Returning to the strategies above, a child might use the gist strategy but might attend to only a few numbers, perhaps 3–4, when comparing two sets of 10 numbers. In such a case, the child might generate a summary not representative of the full dataset, leading to inaccurate conclusions.

Multiple strategies

As suggested above, multiple strategies may be used on the same trial. For example, if a gist strategy did not yield a conclusive answer, a child might use the same strategy again or might use a different strategy (e.g., comparing the highest values). This is consistent with the predictions of the test-operate, test-exit model in which strategies are implemented, evaluated, and, if a goal is not met, additional operations are implemented. In general, multiple strategies tend to be used more frequently when solving difficult problems compared to simpler problems (Carr & Alexeev, 2011; Rittle-Johnson & Siegler, 1999) and the use of multiple strategies in a single trial has been linked to subsequent improvements in reasoning (Fujimura, 2001; Siegler, 2007). As outlined above, the overlapping waves model provides a framework for explaining differences in individual strategy use, fidelity of use, and multiple strategy use.

Present study

The goal of the current study was to identify and evaluate the accuracy of children’s spontaneous data comparison strategies, as inferred from their eye movement behavior (see Table 2). Although the task was similar to that of Morris and Masnick (2015), except with children, the present study advances our understanding not only by extending the sample, but also by conducting a detailed strategy analysis derived from the visual fixation patterns. These strategies illuminate what information is being used and when in real time during the comparison process. To this end, we presented fourth and sixth grade participants with 36 sets of data that varied systematically in the mean ratio, relative variance, and number of observations (details below). We measured eye fixations, accuracy, and confidence (using a 4-point scale), and a brief post-experimental survey of strategy use and knowledge of averaging (i.e., calculating averages from data sets). Fourth and sixth grade children were selected because both have multiple years of experience with place values and the mathematical skills to perform basic calculations (e.g., addition, multiplication). For this reason, all children should exhibit fluency with three-digit numbers. However, these grades differ in their experience with data and probability. The Common Core Math Standards for sixth grade outline basic concepts to be covered, including samples and populations, measures of center, and variance (National Governors Association, 2010). In contrast, the fourth grade standards do not include statistics and probability (e.g., central tendency,
variance). These data were collected after the Common Core had been implemented in each district. In sum, both groups should be equally proficient with individual numbers but might show differences in formal skills for making sense of numbers in sets.

We first coded each child’s patterns of eye movements (fixations, saccades, etc.) for each trial to identify the strategy or strategies used. We then evaluated differences in strategy accuracy by the properties of the data sets (e.g., mean ratio). However, children may use a strategy but implement it with low efficacy; thus, differences in accuracy levels should be compared within the same strategy to investigate this possibility. The identification of differences in strategy use, particularly across ages, would suggest developmental changes in summarizing statistical properties. Another possibility is that children may attempt to summarize the statistical properties but that their attention to nondiagnostic features of set members may skew the resulting representations. For example, if a child focuses the majority of her attention on one number in a set, this number may be more heavily weighted in the summary representation (e.g., underestimating or overestimating an average).

The use of different strategies or ineffective implementation would suggest an emerging understanding of data sensemaking. For example, a child who compares the first number from each set may suggest a belief that each number is equally informative, and that conclusions about sets do not require attention to the full data set. In addition, we analyzed individual variation in strategy use within- and between-trials to investigate how strategies are used to achieve a goal. Of particular interest is whether children used a single strategy or if they demonstrate flexible approaches that suggest a dynamic, iterative approach to making comparisons. Finally, we predict the use of multiple strategies will be more frequent given sets that are more similar than sets that are more distinct. We provide an operational definition for these terms below.

**Method**

**Participants.** A G*Power analysis was conducted before collecting data, indicating that a sample size of 36 per grade would allow for the detection of large effects (.70) with .80 power. Participants were 44 4th grade (mean age = 9.78, SD = .60, 48% female, 79% White, 11% African-American, 9% Asian-American) and 38 6th grade (mean age = 12.14, SD = .44, 53% female, 71% White, 18% African-American, 10% Asian-American) students from three public schools in the Midwestern United States. We will refer to participants by grade level throughout the remainder of the paper. Our study protocol received ethics approval from the Kent State University IRB, **Identifying data comparison strategies in children and adults using eye tracking.** The average household income in the school districts was $34,000, $33,000, and $36,000, respectively. All districts were eligible for free and reduced lunch for 100% of students. Twelve participants (five 4th and seven 6th graders) wore glasses and reported no difficulties seeing the number sets when they removed their glasses, which was necessary for accurate eye-tracking readings.

**Materials.** Participants saw 36 numerical set pairs organized in two columns on either side of a computer screen. Each set pair varied on the following properties: (a) set size: 4, 6, or 8 three-digit numbers per set, (b) mean ratio: a between set ratio of means of either 4:5 (relatively large difference) or 9:10 (relatively small difference), and (c) coefficient of
variation within set of either .10 or .20 (equivalent across sets). For each of the 12 possible combinations of mean ratio, coefficient of variation, and set size, there were three trials with numerical set pairs meeting these specifications, for a total of 36 experimental trials.

Numbers were presented in 42-point Times New Roman font and each column of numbers was centered on each side. Within each number, an extra space was placed between the hundreds, tens, and ones places and 1.5 vertical spacing was used between numbers in each column. A Tobii T-60XL eye tracker was used for data collection. Participants were seated in front of a 17-inch monitor and adjusted position until approximately 70 cm from the monitor (the experimenter’s screen displayed this information in real time to ensure distance and optimal angle precision). We provide more eye tracking details in the Supplementary materials.

Procedure. The eye tracking data were collected by the first author. After being seated in front of the eye tracker, participants were instructed to keep their heads and bodies still during the experiment. A nine-point calibration was performed with the eye tracker. After completing the calibration, participants were told that the session would begin. They were then provided with the following instructions: You will be shown the results from a series of golf drives (a single golf shot to achieve maximum distance). Each slide will show how far a golfer hit a series of balls from one of two tees (LEFT or RIGHT). Your job is to tell me which golfer, on average, hit the ball FARTHER (all drives were measured in feet).1 Participants responded verbally and the experimenter recorded their responses.

Next, participants were given the following instructions about the confidence scale: After you make your choice, I will ask you how sure you are using the scale in front of you. The scale goes from 1 to 4. A 1 indicates that you were NOT SO SURE about which golfer hit the ball farther and a “4” indicates that you were TOTALLY SURE that one golfer hit the ball farther. I will ask you: How sure are you that this golfer hit this ball farther? Please say a number. Participants verbally responded and the experimenter recorded the response.

On each trial, participants first saw a fixation slide in which a + was placed in the center of the screen for 1 second. After one second, participants saw a data slide consisting of the two sets of data (positioned on the left and right sides of the screen) and were asked which golfer hit the ball farther and to rate their confidence in this difference. After they produced both answers, the experimenter advanced the screen to the next fixation slide for 1 second. No participant required prompts during the experimental phase to produce a response. Data sets were presented in blocks by set size and within each block mean ratio and variance were randomized into two counterbalanced orders.

Defining Areas of Interest (AOI). AOIs for eye fixations were defined around stimuli before data collection. AOI’s were defined around the hundreds, tens, and ones columns and around each three-digit number. The number of fixations and duration of fixations occurring within each AOI were automatically recorded using Tobii studio. Coding details are provided in the Supplementary materials.

---

1Alternative instructions were piloted (e.g., Which is the better golfer? The best golfers can hit the ball the farthest. Based on these numbers, which golfer would you choose for your team for an upcoming tournament?). These alternative wordings led to the same response patterns suggesting that the use if the term ‘average’ did not materially alter task demands. We also collected self-report strategy data that is reported in the supplemental materials.
Results

We begin with analyses of accuracy, present the confidence data, analyses of the eye-tracking data, aggregated strategy data, and then analyses of individual strategy use. Comparisons for gender yielded non-significant results, so gender was not considered a factor in the following analyses (see Section S1 in the Supplementary materials for full gender analyses). Grade is considered a between-subjects factor for the following analyses because sixth graders performed better than fourth graders, $t(71.80) = 3.66, p < .001$.

Accuracy. We considered a result to be accurate when a participant chose the set with the higher mean as the answer to which golfer, on average, hit the golf ball farther. The dependent variable was the total number of accurate responses (e.g., choosing the set with the higher mean) of the 36 trials. We conducted a 4-way mixed ANOVA with grade as a between-groups variable (grade: 4th and 6th) x 3 (set size: 4, 6, & 8) x 2 (mean ratio: 4:5 vs. 9:10) x 2 (coefficient of variation: High vs. Low) analysis. The results indicated that accuracy was related to set properties as predicted (see Table 3 for detailed results). We briefly describe the results most relevant to the hypotheses described above and include effect sizes. There were main effects for grade, $F(1, 79) = 86.36, p < .001$ partial $\eta^2 = .99$; mean ratio, $F(1, 79) = 46.97, p < .001$ partial $\eta^2 = .49$; set size, $F(1, 79) = 6.76, p < .001$ partial $\eta^2 = .20$; and variance, $F(1, 79) = 3.23, p < .001$ partial $\eta^2 = .09$ (see Table 3 below for the descriptive statistics and section S2 in the Supplemental materials for full results). Tukey’s HSD indicated that sixth graders ($M = .90, SD = .08$) were more accurate than fourth graders ($M = .76, SD = .12; p < .001$). In addition, children were more accurate for sets with a 4:5 mean ratio ($M = 1.8, SD = .12$) than a closer 9:10 mean ratio ($M = .1.69, SD = .11; p = .01$), set sizes of 4 and 6 ($M = 1.67, SD = .12; p = .001$) compared to sets of 8 ($M = 1.59, SD = .13; p = .03$), and with a smaller coefficient of variation (10% of mean) ($M = 1.79, SD = .11; p < .001$) than larger coefficient of variation (20% of mean) ($M = 1.59, SD = .10; p = .01$). A two-way interaction indicated that children were more accurate when comparing small sets with 4:5 mean ratios than larger sets with 4:5 mean ratios, $F(1, 79) = 72.89, p < .001$, partial $\eta^2 = .21$ (there were no other significant two-way interactions). Finally, there was a three-way interaction between set size, mean ratio, and variance that demonstrated that

| Grade | Accuracy (Max = 1) | Confidence (Max = 4) |
|-------|--------------------|----------------------|
|       | Mean   | SD    | Mean  | SD    |
| Sixth | Overall Mean      | 0.76   | 0.12  | 3.13  | 0.62  |
|       | Set Size = 4      | 0.75   | 0.12  | 3.21  | 0.64  |
|       | Set Size = 6      | 0.76   | 0.15  | 3.14  | 0.67  |
|       | Set Size = 8      | 0.74   | 0.16  | 3.14  | 0.65  |
|       | 4:5 Mean Ratio    | 0.84   | 0.11  | 3.24  | 0.69  |
|       | 9:10 Mean Ratio   | 0.73   | 0.15  | 2.99  | 0.67  |
|       | High Variance     | 0.72   | 0.14  | 3.06  | 0.66  |
|       | Low Variance      | 0.86   | 0.15  | 3.18  | 0.64  |
| Fourth| Overall Mean      | 0.90   | 0.08  | 3.23  | 0.58  |
|       | Set Size = 4      | 0.92   | 0.11  | 3.24  | 0.59  |
|       | Set Size = 6      | 0.93   | 0.10  | 3.25  | 0.61  |
|       | Set Size = 8      | 0.85   | 0.12  | 3.22  | 0.59  |
|       | 4:5 Mean Ratio    | 0.96   | 0.06  | 3.32  | 0.63  |
|       | 9:10 Mean Ratio   | 0.84   | 0.12  | 3.07  | 0.62  |
|       | High Variance     | 0.87   | 0.10  | 3.14  | 0.63  |
|       | Low Variance      | 0.93   | 0.09  | 3.25  | 0.61  |
Making sense of data

Figure 1. Mean accuracy levels by grade and stimulus type. Note: HM = high mean difference, LM = low mean difference, HV = high set variance, LV = low set variance.

children were more accurate with small sets with 4:5 mean ratio and low variance than large sets with 9:10 mean ratio, and high variance, $F(1, 79) = 72.89, p < .001$, partial $\eta^2 = .08$ (see Table 3 and Figure 1 for details).

Confidence. As with the accuracy data, we conducted a factorial ANOVA with three within-subjects factors resulting in a 3 (set size: 4, 6, & 8) x 2 (mean ratio: 4:5 vs. 9:10) x 2 (coefficient of variation: High vs. Low) analysis, with self-reported confidence ratings as the dependent measure (See Table 3 for descriptive data and S2 in the Supplemental materials for full results). There was no main effect of grade or gender and thus the analyses were conducted on all participants’ data as one group. There was a main effect of mean ratio, $F(1, 79) = 72.89, p < .001$, Partial $\eta^2 = .43$, indicating higher confidence given a larger, 4:5 mean ratio than a smaller, 9:10 mean ratio. A main effect of set size was also present, $F(2,78) = 5.34, p = .007$, Partial $\eta^2 = .10$, and a Tukey’s HSD indicated a significant drop in confidence from sets of four to six and no difference between sets of six and eight. Confidence was greater when sets had a lower coefficient of variation (10% of mean) than a higher coefficient of variation (20% of mean), $F(1, 79) = 11.88, p = .014$, Partial $\eta^2 = .21$. As expected, the mean ratio by coefficient of variation interaction was significant. Participants were more confident in their comparisons for sets with 4:5 mean ratios and low variance than sets with 9:10 mean ratios and high variance, $F(1, 79) = 25.64, p < .001$, Partial $\eta^2 = .29$. No other interactions were significant.


**Eye tracking analyses**

We analyzed the eye tracking data in three ways. First, we analyzed the relation between the number of fixations and the properties of data sets. Second, we investigated the number of times children switched their attention between the two data columns as an index of difficulty. Third, we coded eye tracking patterns within each trial to determine strategy use.

*Number of fixations.* A within-subjects ANOVA was used to investigate the number of fixations by data properties. We did not include duration results because they were highly correlated with fixation results. Not surprisingly, the number of fixations increased as set sizes increased, $F(1, 79) = 12.296, \ p < .001$, Partial $\eta^2 = .52$. There was an interaction between set means and variance demonstrating. Specifically, there were more fixations for sets with 9:10 mean ratios and high variance than for sets with 4:5 mean ratios and low variance, $F(2, 79) = 4.8, \ p = .008$, Partial $\eta^2 = .37$. There were no other significant interactions between confidence, mean ratio, and variance. There was also an interaction between set size and mean ratio with more fixations for larger sets compared to smaller sets, only in sets with 9:10 mean ratios, $F(2, 79) = 4.51, \ p = .012$, Partial $\eta^2 = .12$ but no difference for sets with a 4:5 mean ratio. No other interactions were significant.

*Column Switches.* The number of switches between columns was negatively related to mean accuracy, $r = -.23, \ p = .049$, suggesting that making fewer switches between columns was related to better accuracy. The results of a factorial ANOVA indicated that fourth graders had more column switches than sixth graders ($F (1, 77) = 9.55, \ p = .001$, Partial $\eta^2 = .24$) and that the mean ratio by variance interaction was significant, $F (1, 77) = 4.70, \ p = .033$, Partial $\eta^2 = .11$ for both grades (see Figure 2). Tukey’s HSD

---

**Figure 2.** Mean number of column switches by trial type and grade. Note: HM = high mean difference, LM = low mean difference, HV = high set variance, LV = low set variance.
indicated that, across all set sizes, sets with high (i.e., 4:5) mean ratios and low variance (10% of set mean) \((M = 3.39, SD = .47)\) were associated with the fewest switches whereas sets with low (9:10) mean ratios and high variance (20% of set mean) were associated with the most switches \((M = 5.08, SD = .73; p < .001)\). None of the other interactions were significant. The factorial ANOVA also indicates one main effect, that there were fewer switches for sets with a 4:5 mean ratio than 9:10 mean ratio, \(F(1, 77) = 34.40, p < .001\), Partial \(\eta^2 = .43\), whereas there were no differences for coefficient of variation and set size \((F^2's < 1)\).

**Strategy analysis**

Table 4 presents the proportion of strategies used during the task for all participants on all trials. There were 2,664 total trials across all participants and 4.4% of the trials were unable to be coded due to missing data. Two thousand five hundred fifty-seven (2557) strategies were coded and 30% of trials were double coded by two hypothesis-naive coders using the coding criteria described in Table 2. The initial comparison yielded 97% reliability \((\kappa = .83)\). All disagreements were resolved following a discussion with the first author. A chi-square analysis demonstrated a significant difference in the distribution of strategies by grade \(X^2(1, N = 1630) = 67.13, p < .001\) (see Table 4 for proportions and standardized residuals).

The next analysis investigated accuracy by strategy and grade. As above, the analysis compared accuracy by the total number of strategies (2557) independent of participant (see Table 5). An ANOVA demonstrated significant differences in accuracy by strategy and grade, \(F(5,240) = 7.3, p = .001\), Partial \(\eta^2 = .48\). Tukey’s HSD post hoc tests revealed that sixth graders \((M = .90, SD = .08)\) were more accurate than fourth graders \((M = .76, SD = .12; p = .003)\) regardless of strategy. As stated above, we could not conduct meaningful analysis for sixth graders due to the highly uneven distribution of strategies. Fourth graders were somewhat more accurate when using the comparison strategy \((p = .01)\) than any other strategy.

To test the possibility that children might be summarizing the statistical properties that were weighted by attention to specific items (as in the comparison strategy), we examined accuracy on the only five trials in which the first number in each set was not diagnostic of

| Table 4. Proportion of strategy use by grade. |
|---------------------------------------------|
|                | Comparison | Gist | High-Low |
|----------------|------------|------|----------|
| Fourth Graders | .30 (.421) | .45 (−.12) | .25 (.48) |
| Sixth Graders  | .15 (−.439) | .74 (.534) | .11 (−.67) |

Note. Standardized residuals in italics.

| Table 5. Mean accuracy by strategy, for each grade. |
|---------------------------------------------|
|                | Fourth | Sixth |
|----------------|-------|-------|
| Gist           | .72 (.17) | .94 (.13) |
| Comparison     | .81 (.13) | .87 (.11) |
| High-Low       | .74 (.17) | .89 (.12) |
| Overall        | .76 (.16) | .90 (.12) |

Note. Standard Deviation in parentheses.
the set. For example, we looked at cases where the set on the right had the higher average despite the first number in the left set being higher than the first number in the right set. If a child makes comparisons only based on the first two numbers in each set, he or she should answer incorrectly. However, if these features are merely weighted as part of averages, the child should answer correctly. As shown in Table 4, some children did demonstrate the comparison strategy via their eye tracking behavior. The results demonstrated that accuracy for children using the comparison strategy on these trials (32% accuracy) was lower than those using any other strategy (78%) on the same data sets. This finding demonstrates that these children were not summarizing the properties of the entire set while weighting the first item more heavily. Instead, some children were comparing single values from each set while effectively ignoring the set properties. This result provides evidence against the idea that children initially make use of expected strategies for summarizing the statistical properties of data sets.

Why were fourth graders less accurate when using the same strategies?

One interesting result is that fourth graders were less accurate in using the gist strategy than sixth graders. At the level of strategy identification, both ages demonstrated the same pattern of focusing on multiple numbers within the same column, then focusing on multiple numbers within the other column. The difference in accuracy suggests differences in the fidelity of strategy use. To investigate this, we conducted an analysis of two features that children attended to within each strategy. One, we identified the proportion of numbers within each column the child attended to within each trial. This may influence accuracy because attending to a higher proportion of numbers (e.g., attending to all numbers) should be associated with higher accuracy than attending to a smaller proportion of available numbers. This was coded by using the fixation count tool in Tobii Studio to identify the pattern of fixations within the AOIs in each column for each strategy. These fixations were counted to identify if a number in a set had been attended to and then to determine a proportion of attended numbers. For example, in a set of eight numbers, if the AOIs of five numbers were attended to, the proportion for that column would be 0.625.

Two, we identified place values attended to within each column. As mentioned earlier, a focus on hundreds place values will be more diagnostic than a focus on ones place values. This was coded by using the fixation count tool in Tobii Studio to identify fixations within the AOIs around the place value of each number. We totaled the fixations in each place value and divided the total by the number for each place value to determine a proportion for that column. For example, if a participant made 16 fixations within a column, 8 in the hundreds place, 4 in the tens, and 4 in the ones, then this would be coded as .5, .25, and .25, respectively, with 50% of fixations on the hundreds place, and 25% each on the tens and ones. The results of these analyses are displayed in Tables 6 and 7 below. The results

| Table 6. Proportion of numbers fixated upon when using each strategy. |
|---------------------------------------------------------------|
| **Strategy** | **Fourth** | **Sixth** | **t** | **p** |
|---------------|------------|-----------|------|------|
| Gist          | 0.44 (.15) | 0.71 (.11)| 5.43 | .00001* |
| Comparison    | 0.20 (.06) | 0.23 (.06)| 1.44 | .158 |
| High-Low      | 0.59 (.12) | 0.63 (.09)| 1.51 | .144 |
Table 7. Proportion of fixations within AOIs around place values for the gist strategy by grade.

| Place value | Fourth  | Sixth   | t   | p     |
|-------------|---------|---------|-----|-------|
| Hundreds    | 0.76 (.15) | 0.92 (.11) | 5.2 | <.001* |
| Tens        | 0.52 (.12) | 0.33 (.06) | 7.47 | <.001* |
| Ones        | 0.47 (.10) | 0.24 (.04) | 13.4| <.001* |

demonstrate two age-related differences in the fidelity of strategy use. One, fourth graders using the gist strategy attended to significantly fewer numbers than sixth graders using the same strategy. There were no significant differences for other strategies (see Table 6 below).

Two, fourth graders attended to less diagnostic features of the three-digit numbers themselves, specifically, the tens and ones place values, than sixth graders. Sixth graders attended to the hundreds place value, the most diagnostic place value, more than fourth graders (see Table 7). Taken together, these results clarify why fourth graders’ use of the gist strategy resulted in less accurate comparisons. Fourth graders used the gist strategy but attended to fewer numbers and were more likely to attend to less diagnostic number features, specifically the ones place value, than sixth graders, who focused on more numbers in each set and more diagnostic features of the numbers themselves (i.e., hundreds place value). In other words, sixth graders encoded a more comprehensive gist representation, leading to more accurate estimates.

Individual consistency. An important question is the degree to which participants were consistent in their use of a particular strategy. This is important because strategy variability is associated with partial, and often transitional, problem solving approaches (Siegler, 2007). The first analysis measured the degree to which participants used one strategy consistently throughout the task. We coded consistency as the use of one strategy on at least 75% of trials. We then compared the differences between strategy use and grade (See Figure 3). The results indicated that the distribution of strategies significantly differed from chance $X^2 (1, N = 79) = 14.4051, p = .006$. We used standardized residuals to determine significant differences between cells because these values are similar to $z$-scores. The results indicated

![Figure 3. Strategy consistency by grade.](image-url)
that more sixth graders (1.53) than fourth graders (−1.49) were classified as using the gist strategy consistently while more fourth graders (1.99) than sixth graders (−2.09) were classified as using no strategy consistently.

Accuracy was then compared between participants using the gist strategy consistently and those not using any strategy consistently. Results indicated no significant differences in accuracy between sixth graders using the gist strategy consistently (M = .93, SD = .18) and sixth graders using no strategy consistently (M = .88, SD = .22), t(78) = .774, p = .306. Fourth graders coded as not using a strategy consistently (M = .79, SD = .23) were significantly more accurate than fourth graders who consistently used the gist strategy (M = .70, SD = .16), t(78) = 1.88, p = .033. Sixth graders were highly consistent in strategy use across the problem set but fourth graders’ strategy use was much more variable. This may suggest that most sixth graders’ strategy use reflected an understanding of which strategy was most effective and under what conditions it was most effectively implemented. In contrast, the fourth graders who consistently used the gist strategy were accurate most of the time, but still often erred in their conclusions, even more than those who used a variety of strategies. This result is consistent with prior work on children’s knowledge about averages and variability that suggested that children may, at least initially, have a partial understanding of statistical concepts such as averages (Watson & Moritz, 1999). It also suggests the fourth graders’ variation in strategy use may have been an effective approach for sets with relatively large differences but much less effective for less distinct sets.

**Multiple strategy use**

The next analysis investigated whether children were using multiple strategies within a single trial. Multiple strategy use has been documented in many domains (e.g., mathematics, spelling, playing tic-tac-toe; Crowley & Siegler, 1993; Siegler, 2007). The first analysis was to determine the average number of strategies that participants used across trials. Sixth graders used a mean of 1.89 strategies (SD = 1.03) and fourth graders used a mean of 2.61 strategies (SD = 1.04) across the task (t(79) = 3.09, p = .003) with 59% of sixth graders and 90% of fourth graders using more than one strategy on at least one trial.

We next investigated the relation between the use of multiple strategies and problem features. The use of multiple strategies is likely to occur on more difficult problems than simple problems. We have suggested that because mean ratio is the most salient feature, that problems with high mean ratios should be simpler than those with low mean ratios. In addition, based on previous research (Morris & Masnick, 2015), we expected variance to interact with mean ratios in making sets with large mean differences and low variance more easily discriminable than sets with a small mean difference and high variance. To investigate this, we compared the relation between problem features and the use of multiple strategies across ages.

We identified trials on which multiple strategies were used and identified the problem features of each trial. We then calculated the mean number of multiple strategy uses for each problem type. No formal analyses were run for sixth graders due to the small number of observations (25 trials). For fourth graders, a within-subjects ANOVA indicated that the use of multiple strategies differed significantly by problem features. Multiple strategies were
more frequent for sets with low mean differences than for sets with high mean differences, $F(1, 79) = 5.6, p < .014$, Partial $\eta^2 = .14$. There was an interaction between mean ratios and variance, $F(2, 79) = 3.4, p < .037$, Partial $\eta^2 = .12$. A Tukey’s HSD post hoc test demonstrated that there were more strategies used for sets with 9:10 mean ratios and high variance ($M = 2.3, SD = .22; p < .001$) than sets with 4:5 mean ratios and low variance ($M = 1.4, SD = .17; p < .001$). No other differences were statistically significant. We also asked participants to self-report their strategies. These data did not provide additional information beyond the eye tracking data but are presented in the supplementary materials (see section S3).

**Accuracy for averaging**

The second part of the post-trial survey presented four averaging questions. These questions provided number sets and four response options. Participants were asked to select the average of each set. Fourth graders ($M = 1.02, SD = 1.1$) chose correct answers significantly less frequently than sixth graders ($M = 2.49, SD = 1.36$) $t(78) = 5.3, p < .001, d = 1.18$. We then analyzed the extent to which calculation accuracy explained unique variance in comparison accuracy. The results indicated that a small amount of variance in comparison accuracy was explained by calculation accuracy ($r^2 = .126, b = .35, t(78) = 3.35, p < .001$), suggesting that greater knowledge of calculating a mean provided a small, but significant, benefit for the accuracy of informal comparisons. In addition, a small amount of variance in confidence was explained by calculation accuracy ($r^2 = .15, b = .39, t(78) = 4.1, p < .001$) suggesting that greater knowledge of calculating a mean increased confidence in informal comparisons.

**Discussion**

Our experiment investigated how US children make sense of numerical data by investigating the strategies children used to summarize the properties of number sets. Accurately summarizing sets requires the use of effective strategies that direct attention to a sufficient proportion of diagnostic features. Our results extend knowledge of how US children make sense of data and their emerging understanding of how individual numbers are related to each other in sets in three ways. One, the data demonstrate that children initially use a variety of strategies to summarize the statistical properties of data sets. This pattern suggests that the use of expected strategies with fidelity is the result of a developmental and educational progression. Two, our results indicated that the statistical properties of data sets were a strong predictor of child accuracy and confidence suggesting that, at least for highly distinct sets, children’s initial strategies may be sufficient to compare sets accurately. Three, we identified a series of strategies US children used to compare data sets and coded the consistency of their use. Our strategy coding was derived from eye tracking data during task trials and provides a measure of process tracing, or identifying the elements to which children attend in real-time as they solve problems (Schulte-Mecklenbeck et al., 2017). Our results demonstrated that several of the younger children used multiple strategies, often within the same trial, to make sense of the data.
Adults are quite adept at summarizing the statistical properties of sets (Whitney & Yamanashi Leib, 2018). Summarizing these properties from complex information provides useful information about the environment and reduces the overall processing burden (Alvarez, 2011). Although children can summarize statistical properties in some conditions (e.g., the average size of object sets, Sweeny et al., 2015), they appear to have difficulty with others (e.g., identifying the center from a dot array, Jones & Dekker, 2018). Regardless of context, children need to attend to relevant information in order to create accurate summaries. Our results add to this literature by describing and evaluating the strategies children use to summarize and compare data sets. Sixth graders were highly accurate in their comparisons and consistent in their strategy use. Specifically, nearly all used a gist strategy in which they attended to the most diagnostic features of the entire (or nearly entire) set from which they could summarize an approximate average for comparisons. In contrast, fourth graders were less accurate and less consistent in their strategy use than the older children. Many of the strategies they used were characterized by attention to part of a set or to one feature of the set properties (e.g., highest value). This suggests that younger children may not yet understand that summarizing set properties necessitates attending to the entire data set (similar to the results of Jones & Dekker, 2018). At the same time, younger children who used comparisons and a variety of strategies tended to be more accurate than those who used the gist strategy most consistently.

Why might children vary their strategies more and be able to summarize the statistical properties from some type of stimuli but not others? It may be that the stimuli used by Sweeny et al. (2015) provided supports that helped children understand the relations between elements. Specifically, in the Sweeny et al. (2015) experiment, children saw four oranges in one tree and four oranges in a different tree. The object sets were in close physical proximity and were conceptually and contextually related. These cues may have provided support that the objects were meaningfully related. In contrast, though the number sets used in the present experiment shared perceptual similarity, their relation to the concepts being measured may have been less familiar, weakening the cues that guide attention to relevant features and the summary of statistical properties. Further, the variation in the sets may have led children to differing strategies on different sets. Future research should investigate this potential limitation by investigating children’s performance across different types of stimuli to measure the role of contextual support in performance.

The statistical properties of data sets emerge from the numbers in the set but are not present in any individual number. The results indicated that aggregated accuracy and confidence were closely related to the statistical properties of data sets, particularly when children directed their attention to the relevant features of the numbers in the set (e.g., using a gist strategy). The ratio of means was the strongest predictor of accuracy and confidence, indicating that relatively large differences between sets were easier to detect than smaller differences, regardless of strategy. Accuracy and confidence decreased as set sizes increased, suggesting that representing properties was easier for smaller sets. Variance was associated with main effects for accuracy and confidence and consistently interacted with mean ratios and set sizes by magnifying their effects, consistent with previous findings (e.g., Obrecht et al., 2007).

Eye fixation data provided converging evidence that set properties had a strong influence on performance. For example, participants made a greater number of fixations on numbers in the sets with small mean differences compared to sets with larger mean differences. This
suggests that, in the aggregate, more attention was recruited to distinguish less distinct sets (e.g., sets with smaller mean differences and high coefficient of variation). Finally, participants made more column switches when comparing less distinct sets and accuracy decreased as the number of switches increased. These results suggest that less distinct sets cued the search for additional information during comparisons. A potential alternative explanation is that the increase in the number of column switches might indicate information processing overload. However, evidence from previous research suggests that information processing overload is associated with fewer fixations and gaze aversion, rather than increased fixations (Doherty-Sneddon & Phelps, 2005).

Younger children were less consistent in their strategy use than older children. Specifically, while over half of sixth graders consistently used the gist strategy (i.e., used a single strategy on >75% of trials), the majority of fourth graders used no strategy consistently. Previous research has suggested that the use of multiple strategies may indicate a transitional point in the learning process (Siegler, 2005). The strategies used by fourth graders also suggested a focus on individual numbers, rather than attention to set properties. For example, most fourth graders used a comparison or high-low strategy on at least one trial, which demonstrates more attention to individual number than set properties. Comparing only the highest value in each set indicates that such values carry information about the relations between numbers in a set, while focusing on the first value in the set suggests a belief that any value is equally diagnostic. Either strategy ignores other potentially relevant set information such as means and variance. Finally, even when younger children used potentially more effective strategies like the gist strategy, they often used them ineffectively. The data demonstrate that this ineffective use is related to attending to a smaller proportion of numbers and non-diagnostic number features (e.g., more attention to the ones place value than the hundreds place value). At the same time, fourth graders who used more varied strategies were more accurate than those who used primarily a gist strategy. This may be due to properties of the sets themselves – when the comparison strategy leads to an accurate answer because the first values are lined up with the means, this may have been considered a more efficient strategy, and it led to accurate answers with many of the datasets.

Children also used multiple strategies within a single trial. The use of multiple strategies fits the overlapping waves theory (Siegler, 2016) in which children’s thinking is variable, children may use multiple strategies, and an initial approach might not yield a definitive result. In the absence of an immediate solution, reasoners might try the same approach again or might change their approaches. Our results provide evidence that fourth graders were more likely to use multiple strategies within a trial than sixth graders. One possible explanation is that younger children were not using strategies with fidelity (Miller, 2000; Waters, 2000). For example, a child might use the gist strategy but might attend to only a few numbers rather than the entire set. In such a case, he or she might not see a clear difference between sets because several of the numbers in each set might overlap. Our data support this in that sixth graders made more accurate comparisons than fourth graders using the same strategies. At the same time, overall the fourth graders did better when using multiple strategies than when using just the gist strategy, in part likely because of their lack of fidelity to the gist strategy use. The results demonstrated that younger children attended
to fewer numbers than older children when using the gist strategy and attended to less salient features (e.g., looking at the ones place value). These factors would result in summaries that were less accurate and could lead to erroneous comparisons.

The use of multiple strategies was related to the properties of the number sets. Multiple strategies were more likely to be used within trials with low mean differences compared to trials with high mean differences. There was an interaction between mean difference and variance in that multiple strategies were most likely to be used for sets with low mean difference and high variance. Further, we found a developmental difference, in that fourth graders used more strategies overall and more strategies within a single trial than sixth graders.

The results provide support for the idea that summarizing the statistical properties of data is the result of a developmental and educational progression. The evidence suggests that fourth graders were less likely to attend to salient features and were more likely to use of multiple strategies across the set of trials, than sixth graders. The latter reflects a child trying out different approaches for a novel problem and is consistent with the overlapping waves model (Siegler, 2016), in which children accrue data from strategy use that allows them to refine the likelihood of future strategy use. As demonstrated in the results on individual numbers and set properties, sixth graders were more likely to attend to diagnostic features of both (e.g., attending to most, if not all numbers in a set). This helps to explain why sixth graders used the gist strategy consistently while fourth graders used a variety of strategies across and within trials. The converging evidence from mean calculation accuracy supports this suggestion in that children with experience calculating the statistical properties of data sets may have more knowledge about the diagnostic properties of numbers in sets.

Finally, the results add to our understanding of how children become more adept at summarizing complex information. Our results suggest that summarizing complex information required visual attention to a sufficient amount of diagnostic information to represent set properties accurately. The focus on identifying and measuring changes in strategy use helps us understand, at a fine grain size, children’s attentional patterns as they summarize complex information. Future research, using similar eye tracking measures, could help identify strategies children use to create ensembles.

The results suggest two educational implications. One, because young children enter classrooms with limited knowledge about data, providing experiences that add to their background knowledge may improve learning outcomes. Allowing children to explore data before instruction may be beneficial because it may help focus their attention on relevant features, which prepares children to learn (Schwartz & Martin, 2004). This may be especially useful when exploration provides comparisons between sets to highlight relevant statistical properties such as variability within sets. Two, it may be easier for children to detect the relevant properties of data from graphical representations than only through the presentation of numerical information (Garfield & Ben-Zvi, 2007). Properties of data (e.g., variance) may be more salient in a graphical representation than raw numerical representations (Schwartz & Martin, 2004). For example, college students learned more about probability distributions when they played a video game that provided dynamic, graphical information about distributions than when they read information about such distributions (Arena & Schwartz, 2014), and a similar effect may occur with children.
One limitation of this study is that we measured only one component of data sensemaking, comparisons. The strategies used to perform comparisons may be dissimilar to those used for other tasks, and thus caution should be used when generalizing to similar tasks. Other possible options for measuring data sensemaking, including generating approximate means and generating predictions from data, should be investigated in future research to determine the extent to which the strategies identified in our results are indicative of more general data sensemaking skills. A second limitation is that no data were collected to measure other, potentially related, numerical and mathematical skills. One example is the accuracy of children’s number representations. As suggested in the introduction, summarization may occur by aggregating multiple approximate number representations. However, the accuracy of summary values is likely dependent on the accuracy of the underlying number representations. When underlying representations are highly inaccurate, summaries of these values are also inaccurate (Morris, Todaro, Arner, & Roche, 2022). Thus, future research should measure the accuracy of individual representations. In addition, more robust assessments of children’s mathematical abilities might be useful in determining math skills and number sense that may contribute to data sense. A third potential limitation is the use of eye tracking to determine strategies. Although the fixation patterns themselves constitute objective data, the interpretation of the patterns is not objective. Future research that makes use of fixation patterns would be stronger with converging measures that bolster their interpretation by excluding potential alternative explanations. That said, the use of fixation patterns in this context was useful for identifying strategies and may be a useful approach for future research.

Conclusions

Children are driven to make sense of data. One way that adults and children make sense of data is by quickly summarizing the statistical properties from data sets. Our results suggest that while young children use many strategies to make sense of data, ineffective strategies are sometimes used and effective strategies are sometimes used with low fidelity. As a result, children may make sense of data by focusing on part of a set (e.g., individual numbers) or values that provide a limited characterization of the set (e.g., highest value). Overall, children were more accurate, more confident in their evaluations, and produced fewer fixations as sets became more distinctive, that is, sets with large ratios of mean difference, low coefficients of variation, and small set sizes. We identified children’s data strategies from their eye tracking patterns. Our results demonstrated that most sixth graders consistently used a strategy characterized by attending to the most diagnostic properties of the entire set. Fourth graders used more strategies across trials, were less likely to use strategies with fidelity, and were more likely to use multiple strategies within a single trial, suggesting that fourth graders are acquiring the foundational intuitions about the emergent properties of numbers in sets. These results suggest that a clearer understanding of children’s strategies may improve our understanding of the cognitive processes underlying making sense of data and may suggest efficacious educational interventions. The results also extend our understanding of how improvements in children’s strategy use help them improve their summaries of the abundant information in their environments. Finally, the results suggest possible
targets for educational interventions that support children’s attention to relevant problem features as well as the acquisition and selection of effective data strategies.

**Acknowledgments**

The data that support the findings of this study are available from the corresponding author, BJM, upon reasonable request.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**References**

Alibali, M. W. (1999). How children change their minds: Strategy change can be gradual or abrupt. *Developmental Psychology, 35*(1), 127. doi:10.1037/0012-1649.35.1.127

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131. doi:10.1016/j.tics.2011.01.003

Arena, D. A., & Schwartz, D. L. (2014). Experience and explanation: Using videogames to prepare students for formal instruction in statistics. *Journal of Science Education and Technology, 23*(4), 538–548. doi:10.1007/s10956-013-9483-3

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162. doi:10.1111/1467-9280.00327

Bjorklund, D. F., Coyle, T. R., & Gaultney, J. F. (1992). Developmental differences in the acquisition and maintenance of an organizational strategy: Evidence for the utilization deficiency hypothesis. *Journal of Experimental Child Psychology, 53*(3), 434–448.

Bonny, J. W., & Lourenco, S. F. (2013). The approximate number system and its relation to early math achievement: Evidence from the preschool years. *Journal of Experimental Child Psychology, 114*(3), 375–388.

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review, 24*(4), 396–439. doi:10.1016/j.dr.2004.08.005

Buckley, P. B., & Gillman, C. B. (1974). Comparisons of digits and dot patterns. *Journal of Experimental Psychology, 103*(6), 1131–1136.

Carr, M., & Alexeev, N. (2011). Fluency, accuracy, and gender predict developmental trajectories of arithmetic strategies. *Journal of Educational Psychology, 103*(3), 617. doi:10.1037/a0023864

Chen, Z., & Klahr, D. (2008). Remote transfer of scientific-reasoning and problem-solving strategies in children. *Advances in Child Development and Behavior, 36*, 419–470.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393–404. doi:10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*(7), 891–900. doi:10.1016/j.visres.2004.10.004

Cordes, S., & Brannon, E. M. (2009). Crossing the divide: Infants discriminate small from large numerosities. *Developmental Psychology, 45*(6), 1583–1594.

Crowley, K., & Siegler, R. S. (1993). Flexible strategy use in young children’s tic-tac-toe. *Cognitive Science, 17*(4), 531–561. doi:10.1207/s15516709cog1704_3

De Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics, 70*(5), 789–794. doi:10.3758/PP.70.5.789

Dehaene, S. (2009). Origins of mathematical intuitions: The case of arithmetic. *Annals of the New York Academy of Sciences, 1156*(1), 232–259. doi:10.1111/j.1749-6632.2009.04469.x

Doherty-Snaddon, G., & Phelps, F. G. (2005). Gaze aversion: A response to cognitive or social difficulty?. *Memory & Cognition, 33*(4), 727–733.
Fazio, L. K., DeWolf, M., &Siegel, R. S. (2016). Strategy use and strategy choice in fraction magnitude comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(1), 1. doi:10.1037/xlm0000153

Fujimura, N. (2001). Facilitating children’s proportional reasoning: A model of reasoning processes and effects of intervention on strategy change. *Journal of Educational Psychology, 93*(3), 589. doi:10.1037/0022-0663.93.3.589

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, 75(3), 372–396.

Gaultney, J. F., Kipp, K., & Kirk, G. (2004). Utilization deficiency and working memory capacity in adult memory performance: Not just for children anymore. *Cognitive Development, 20*(2), 205–213.

Gibson, E. J., 1969. *Principles of perceptual learning and development.* New York: Appleton-Century-Crofts.

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science, 2*(2), 265–284.

Goldstone, R. L., Marghetis, T., Weitnauer, E., Ottmar, E. R., & Landy, D. (2017). Adapting perception, action, and technology for mathematical reasoning. *Current Directions in Psychological Science, 26*(5), 434–441.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*(3), 337–364. doi:10.1207/s15326985ep2703_5

Hayhoe, M., &Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Science, 9*(4), 188–194. doi:10.1016/j.tics.2005.02.009

Heydt, R., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin, 138*(6), 1172. doi:10.1037/a0029333

Irwin, F. W., & Smith, W. A. S. (1957). Value, cost, and information as determiners of decision. *Journal of Experimental Psychology, 54*(3), 229–232.

Jones, P. R., & Dekker, T. M. (2018). The development of perceptual averaging: Learning what to do, not just how to do it. *Developmental Science, 21*(3), e12584. doi:10.1111/desc.12584

Kuhn, D. (2013). Reasoning. In P. D. Zelazo (Ed.), *Oxford handbook of developmental psychology (Vol. 1): Body and mind* (pp. 744–764). Oxford, UK.: Oxford University Press.

Lathrop, R. G. (1967). Perceived variability. *Journal of Experimental Psychology, 73*(4p1), 498–502.

Masnick, A. M., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development, 79*(4), 1032–1048. doi:10.1111/j.1467-8624.2008.01174.x

Masnick, A. M., & Morris, B. J. (2022). Data reasoning as a scientific reasoning competency. *Education Sciences, 12*(2), 71. doi:10.3390/eduscience12020071

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior.* New York: Henry Holt and Co. doi:10.1037/10039-000

Miller, P. H. (2000). How best to utilize a deficiency. *Child Development, 71*(4), 1013–1017. doi:10.1111/1467-8624.00205

Morris, B. J., &Sloutsky, V. (2001). Children’s solutions of logical versus empirical problems: What’s missing and what does? *Cognitive Development, 16*(4), 907–928. doi:10.1016/S0885-2041(01)00070-3

Morris, B. J., & Masnick, A. M. (2015). Comparing data sets: Implicit summaries of the statistical properties of number sets. *Cognitive Science, 39*(1), 156–170. doi:10.1011/cogs.12141

Morris, B. J., Todaro, R., Arner, T., & Roche, J. (2022). Do number representations influence the accuracy of children’s numerical predictions? *Frontiers in Developmental Psychology, 13*. doi:10.3389/fpsyg.2022.874230

Moyer, R. S., &Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215*(5109), 1519–1520.

National Governors Association. (2010). *Common core state standards.* Washington, DC.
Nuerk, H. C., Weger, U., & Willmes, K. (2001). Decade breaks in the mental number line? Putting the tens and units back in different bins. *Cognition, 82*(1), B25–33. doi:10.1016/s0010-0277(01)00142-1

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive t tests: Lay use of statistical information. *Psychonomic Bulletin & Review, 14*(6), 1147–1152. doi:10.3758/BF03193104

Obrecht, N. A. (2019). Sample size weighting follows a curvilinear function. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(4), 614. doi:10.1037/xm0000615

Park, J., & Sterbinsky, J. (2015). The approximate number system acuity redefined: A diffusion model approach. *Frontiers in Psychology, 6.*

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*(1), 29–46. doi:10.1037/h0024722

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*(1), 29–46.

Petrosino, A. J., Lehrer, R., & Schaub, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning, 5*(2–3), 131–156. doi:10.1080/10986065.2003.9679997

Rhodes, G., Neumann, M., Ewing, L., Bank, S., Read, A., Engfors, L. M., . . . Palermo, R. (2016). Ensemble coding of faces occurs in children and develops dissociably from coding of individual faces. *Developmental Science, 21*(2), e12540. doi:10.1111/desc.12540

Rittle-Johnson, B., & Siegler, R. S. (1999). Learning to spell: Variability, choice, and change in children’s strategy use. *Child Development, 70*(2), 332–348. doi:10.1111/1467-8624.00025

Rothkopf, C. A., & Pelz, J. B. (2004) Head movement estimation for wearable eye tracker. In *Proceedings ACM SIGCHI: Eye Tracking Research and Applications* (pp.123–130). ACM Press.

Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., Goldstein, D. G., Russo, J. E., Sullivan, N. J., & Willemsen, M. C. (2017). Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science, 26*(3), 442–450. doi:10.1177/0963721417708229

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129–184. doi:10.1207/s1532690xci2202_1

Siegler, R. S. (2005). Children’s learning. *American Psychologist, 60*(8), 769. doi:10.1037/0003-066X.60.8.769

Siegler, R. S. (2007). Cognitive variability. *Developmental Science, 10*(1), 104–109. doi:10.1111/j.1467-7687.2007.00571.x

Siegler, R. S. (2016). Continuity and change in the field of cognitive development and in the perspectives of one cognitive developmentalist. *Child Development Perspectives, 10*(2), 128–133. doi:10.1111/cdep.12173

Sprute, L., & Temple, E. (2011). Representations of fractions: Evidence for accessing the whole magnitude in adults. *Mind, Brain, and Education, 5*(1), 42–47.

Sweeney, T. D., Wurmhansn, N., Gopnik, A., & Whitney, D. (2015). Ensemple perception of size in 4–5-year-old children. *Developmental Science, 18*(4), 556–568. doi:10.1111/desc.12239

Thevenot, C., Barrouillet, P., Castel, C., & Uittenhove, K. (2016). Ten-year-old children strategies in mental addition: A counting model account. *Cognition, 146*, 48–57. doi:10.1016/j.cognition.2015.09.003

Trumpower, D. L. (2013). Formative use of intuitive analysis of variance. *Mathematical Thinking and Learning, 15*(4), 291–313.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin, 138*(6), 1172–1217.

Waters, H. S. (2000). Memory strategy development: Do we need yet another deficiency? *Child Development, 71*(4), 1004–1012. doi:10.1111/1467-8624.00204

Watson, J. M., & Moritz, J. B. (1998). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145–168. doi:10.1023/A:1003594832397

Watson, J. M., & Moritz, J. B. (1999). The development of concepts of average. *Focus on Learning Problems in Mathematics, 21*(4), 15–39.
Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology, 69*(1), 105–129. doi:10.1146/annurev-psych-010416-044232

Young, A. G., Shtulman, A., Pammi, V. S. C., & Dutt, V. (2020). How children’s cognitive reflection shapes their science understanding. *Frontiers in Psychology, 11*, 11. doi:10.3389/fpsyg.2020.00011