COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2014) 1–15

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12141

Comparing Data Sets: Implicit Summaries of the Statistical Properties of Number Sets

Bradley J. Morris,[a] Amy M. Masnick[b]

[a]*Department of Lifespan Development and Educational Science, Kent State University*
[b]*Psychology Department, Hofstra University*

Received 21 November 2011; received in revised form 25 July 2013; accepted 19 November 2013

Abstract

Comparing datasets, that is, sets of numbers in context, is a critical skill in higher order cognition. Although much is known about how people compare single numbers, little is known about how number sets are represented and compared. We investigated how subjects compared datasets that varied in their statistical properties, including ratio of means, coefficient of variation, and number of observations, by measuring eye fixations, accuracy, and confidence when assessing differences between number sets. Results indicated that participants implicitly create and compare approximate summary values that include information about mean and variance, with no evidence of explicit calculation. Accuracy and confidence increased, while the number of fixations decreased as sets became more distinct (i.e., as mean ratios increase and variance decreases), demonstrating that the statistical properties of datasets were highly related to comparisons. The discussion includes a model proposing how reasoners summarize and compare datasets within the architecture for approximate number representation.

Keywords: Number processing; Intuitive statistics; Eye tracking

1. Introduction

Comparing datasets is a critical skill in mathematics, science, and everyday life. People regularly are faced with decisions to compare the health efficacy of multiple treatments, the relative costs of goods and services, and the best choices for investments. All can be considered informal comparisons of data, or numbers in context. For example, how does a consumer determine that a particular store has lower prices than another store? In a comparison of ten products, a shopper finds that seven of these ten are more expensive at

Correspondence should be sent to Bradley J. Morris, Department of Lifespan Development and Educational Science, 412A White Hall, Kent State University, Kent. E-mail: bmorri20@kent.edu

Store A than Store B. Although a statistically inclined shopper could in theory perform a formal comparison such as a *t* test, it seems more likely that she would rely on an informal comparison.

Although there has been a considerable interest in the mechanisms underlying number comparison (see Dehaene, 2009 for a review), there has been little research investigating how people compare *sets* of numbers, despite evidence of its value in statistics and calls for increased focus on teaching students to compare groups of data (Konold & Pollatsek, 2002). In the current article, we investigate the comparison of number sets and propose a model of how number sets are represented and compared that builds upon our knowledge of number representation and comparison.

1.1. Number comparison

Numbers are represented both as approximate magnitudes with error variance and as verbal categories (i.e., exact representations; Dehaene, 2009). That is, given the number "12," we represent both the verbal category "twelve" (an exact value) and an activation on a logarithmically scaled mental number line that peaks around 12 in the approximate number system, and that has constant variability[1] (Dehaene, 2009; Feigenson, Dehaene, & Spelke, 2004; Opfer & Siegler, 2012). Under this model of logarithmic representation with constant variability, differences between single-digit numbers are detected more quickly and accurately as the *ratio* of the difference between numbers increases (Dehaene, 2009). For example, when asked to find the larger number, reaction times are faster and evaluations are more accurate when comparing 3 and 9 (1:3 ratio of numbers) than when comparing 9 and 10 (9:10 ratio of numbers). This *distance effect* is evidence that numerical quantities are compared using approximate representations. Evidence for a distance effect is consistent across presentation formats (e.g., dots, Arabic numbers, fractions; Buckley & Gillman, 1974; Dehaene, 2001; Moyer & Landauer, 1967; Sprute & Temple, 2011) and across age of participants (Feigenson et al., 2004).

Comparisons between multi-digit numbers (e.g., 63 and 72) also demonstrate a distance effect (Dehaene, Dupoux, & Mehler, 1990; Korvorst & Damian, 2008; Nuerk, Weger, & Willmes, 2001; Nuerk, Kaufmann, Zoppoth, & Willmes, 2004). However, two- and three-digit numbers produce distance effects for each place value unit (i.e., tens versus ones; Korvorst & Damian, 2008; Nuerk et al., 2001). Although the highly accurate and quick responses suggest a magnitude representation, the effect of position of specific values (e.g., hundreds place) suggests that category information about place values was more diagnostic of true differences (Korvorst & Damian, 2008). Eye tracking results demonstrated that there were more fixations overall when numbers were closer together and when there was incompatibility between places (e.g., 47 vs. 51, a comparison of two-two-digit numbers, in which one-two-digit number has a larger number in the tens column, but the other one has a larger number in the ones column; Moeller, Fischer, Nuerk, & Willmes, 2009), suggesting that more fixations increase the amount of diagnostic information about each number.

1.2. Comparing number sets

We suggest that when comparing number sets, reasoners create approximate summaries of the statistical properties of these sets, including unique characteristics of sets unavailable in individual number comparisons, such as means and variances (Masnick & Morris, 2008). Recall that approximate numbers are activation functions in which activation levels are highest at the position on a logarithmically scaled number line corresponding to the number being represented and decrease as the distance from the number increases (e.g., Dehaene, 2009; Grossberg & Repin, 2003). It is possible that summaries emerge because the activation of multiple approximate magnitudes for each member of the set results in a summary activation area (see Fig. 1). In other words, if each number is represented as an activation around a range of values, an approximation of relative means could emerge from the degree of overlap between two or more values and approximate variance could emerge from spread of values within the set. Sets with large mean differences and low variance would constitute *high-contrast sets,* in that they would create summary activation areas that would be far apart on a number line. Conversely, sets with low mean differences and high variance would constitute *low-contrast sets* that would be close together on a number line and may overlap. If the summary representations are compared like single number representations, then a distance effect for sets would be expected, and high-contrast sets would be associated with faster, more accurate comparisons and low-contrast sets with slower, less accurate comparisons.

There is evidence of summarizing (i.e., averaging over) sets in other domains (Ariely, 2001; Chong & Treisman, 2003, 2005). When presented with sets of dots of different sizes, adult participants erroneously "recalled" average-sized circles values not in the original array more frequently than actual members of the original set (Ariely, 2001; Chong & Treisman, 2003, 2005). Similarly, adults "summarized" sets of line segments and consumer information showing sensitivity to means and variance (Obrecht, Chapman, & Gelman, 2007; Trumpower & Fellus, 2008; Vickers, Burt, Smith, & Brown, 1985). Van Opstal, de Lange, and Dehaene (2011) demonstrated that participants could approximate means and sums for small number sets. In each case, these summaries appeared to arise without deliberation.

1.3. Current experiment

We asked participants to compare sets of data that varied systematically in the mean ratio, relative variance, and number of observations. We recorded eye fixations, accuracy, and confidence in comparisons. Eye tracking studies have demonstrated that longer fixations and a larger number of fixations are often related to difficulty in solving problems (Reichle, Pollatsek, Fisher, & Rayner, 1998). For example, when reading technical passages, eyes frequently move back to text that has been previously observed, suggesting scanning for additional information related to goal achievement (e.g., comprehension; Rayner & Pollatsek, 1989). Recent work suggests a similar effect in comparing the sizes of two sets of dots: the ratio of set sizes affected eye fixation duration (Libertus & Libertus, 2011).

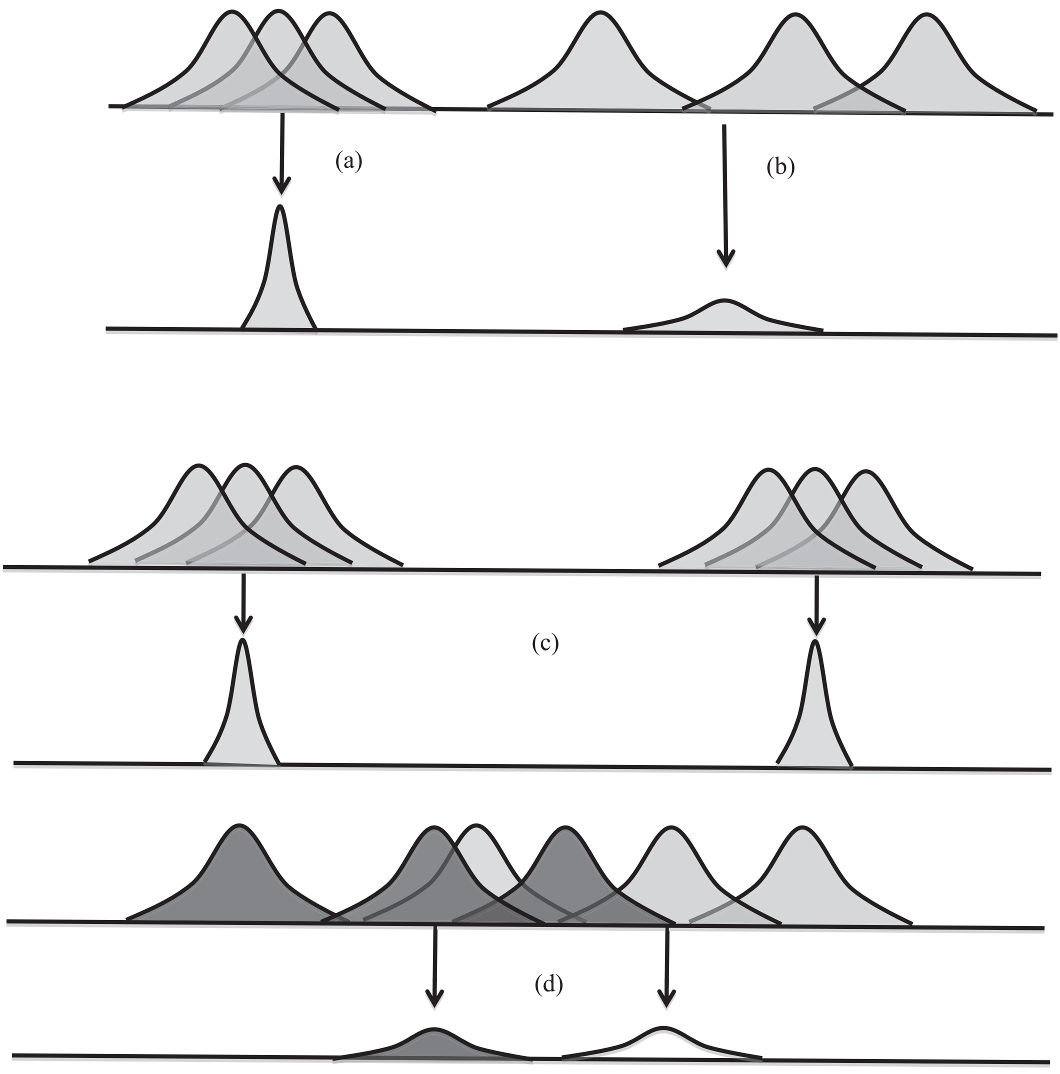

Fig. 1. Number representation. Note that the *x*-axis is log scaled following Dehaene (2009). Panels a and b display two sets of two numbers represented logarithmically with constant variability (above) and summary values represented as secondary activations (below). The numbers in set (a) are close together and result in a summary value with a high peak (approximate mean) and narrow activation area (approximate variance). The numbers in set (b) are farther apart than those in set (a) and result in a summary value with a low peak (approximate mean) and wide activation area (approximate variance). Figure (c) displays a dataset with high ratio of means and low variance—that is, high-contrast sets. Figure (d) displays a dataset with low ratio of means and high variance—that is, low-contrast sets. If difference is detected by comparing distance, then the sets in (c) should be compared more accurately than the sets in (d).

The goal of this experiment was to investigate how number sets are compared using eye fixations, accuracy, and confidence measures. We investigated three questions. One, are features of datasets such as mean and variance related to the accuracy, confidence, and number of fixations during comparisons? Two, is there evidence that participants look at only part of each number instead of the whole number? Three, are participants more likely to switch visual attention between low-contrast sets as compared to high-contrast sets?

2. Method

2.1. Participants

Participants were 24 university students (21 female), who received course credit for their participation. Eight participants were wearing either contact lenses or glasses. Those wearing glasses removed their glasses and reported no difficulties seeing the number sets.

2.2. Materials

Participants saw 36 numerical sets (3 of each type) with the following properties: (a) either 4 or 8 observations per set, (b) ratio of means of either 2:3, 4:5 or 9:10, and (c) coefficient of variation in either .10 or .20 of the mean. All numbers in the datasets were 3-digit numbers. See Table 1 for examples. For each of the 12 possible combinations of mean difference, coefficient of variation, and sample size, there were three trials with datasets meeting these specifications. Thus, there were a total of 36 experimental trials in which these characteristics were varied. Numbers were presented in 42-point Times New Roman font and each column of numbers was centered within two columns in a Power Point slide. Within each number, an extra space was placed between the hundreds, tens, and ones places and 1.5 spacing was used between numbers in each column.

Table 1
Example dataset

LEFT	RIGHT
699	911
820	777
660	733
781	868

Note. These datasets are low-contrast sets; they have means in the ratio 9:10, with a low coefficient of variation, and of course four numbers per set. We computed the coefficient of variation as a standard deviation that was either .10 or .20 the value of each column's mean. Because the two column means were not identical, the variances were not identical. However, a follow-up study with nine subjects in which the variances were identical in each column (the average of the .10 or the .20 coefficients of variation) led to the same patterns of results, in which the participants responded to mean ratio and variance in their assessment of confidence, and the same main effects held for accuracy, confidence, and fixations.

2.3. Procedure

A Tobii T-60 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 (monitor provided this information in real time to ensure distance precision). A nine-point calibration was performed and no participants required recalibration. After completing the calibration, participants were instructed that the session would begin. Participants were given 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).* Alternative 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 response patterns the same as those reported below.

Next, participants were given information 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.*

On each trial, participants first saw a fixation slide in which a + was placed in the center of the screen for 1 s. Then participants saw a data slide consisting of two sets of data (marked LEFT and RIGHT and positioned on the left or right side of the screen). They were then asked to determine which golfer, on average, hit the ball farther and how confident they were in this difference (using the scale positioned in front of them). Datasets were presented in blocks by sample size (4 or 8 observations) and each block was presented in one of two counterbalanced orders. Thus, the overall presentation of blocks was randomized into one of the four possible orders of data presentation.

2.4. Defining areas of interest

Areas of interests (AOIs) were defined around stimuli before data collection. An AOI was 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 was automatically recorded.

3. Results

3.1. Accuracy

Participants were asked to state which column had data from a golfer who hit a golf ball farther. Responses were considered accurate when participants chose the column with the higher mean (95.5% accurate across all trials). A 3-way ANOVA with mean ratio (2:3; 4:5; 9:10), coefficient of variation (.10 of mean; .20 of mean), and sample size (4, 8) as repeated measures independent variables was run, using accuracy as the dependent measure (0-3 possible correct answers for each combination of these factors). The overall accuracy rate was close to ceiling. The only effect was a main effect for mean ratio, with lowest accuracy for 9:10 ratio ($M = 2.79$, $SD = .23$), and higher accuracy for 4:5 ($M = 2.90$, $SD = .18$) and 2:3 mean ratios ($M = 2.97$, $SD = .11$), F (2, 22) = 10.76, $p = .001$, partial $\eta^2 = .50$. Post hoc paired t tests, indicate a significant difference between accuracy for 9:10 ratio and the two other ratios (compared with 2:3 ratio: $t(23) = 4.7$, $p < .001$;4:5 ratio t (23) = 3.32, $p = .003$, respectively). The difference between the accuracy in the 2:3 ratio and the 4:5 ratio trials was not significant No other main effects and none of the interaction effects were significant. Inaccurate trials (39 out of 864) were eliminated from further analyses.[2]

3.2. *Confidence*

In addition to accuracy, participants noted their confidence in their judgments of which golfer was better on a four-point scale from 1 (not sure) to 4 (totally sure). A 3-way ANOVA similar to the one described in the previous paragraph was run with confidence as the dependent measure. There was a main effect of mean ratio; participants had more confidence on the 4-point scale when the means were in a 2:3 ratio ($M = 3.56$, $SD = .24$), than in 4:5 ratio ($M = 3.30$, $SD = .32$), or in 9:10 ratio ($M = 2.49$, $SD = .29$), F (2, 22) = 210.19, $p < .001$, partial $\eta^2 = .95$. Post hoc paired t tests indicate there were significant differences in confidence between all levels (all ps < .001). A main effect of coefficient of variation showed that participants were more confident when the coefficient of variation was .10 ($M = 3.28$, $SD = .29$) than when it was .20 ($M = 2.94$, $SD = .27$), F (1, 23) = 50.03, $p < .001$, partial $\eta^2 = .69$. Finally, there was an interaction between mean ratio and coefficient of variation, such that when the means were very different (high contrast), there was less difference in confidence based on coefficient of variation, F (2, 23) = 9.25, $p = .001$, partial $\eta^2 = .46$. No other main effects or interactions approached significance.

3.3. *Eye fixation counts*

We looked at the behavioral measure of eye fixation counts to explore the pattern of where people look for information in this task (see Fig. 2a and b). In a 5-way ANOVA, mean ratio (2:3, 4:5, 9:10), coefficient of variation (.10 of mean, .20 of mean), sample size (4, 8), side of screen data were presented on (left, right) and column (hundreds, tens, ones) were all repeated measures independent variables, and number of fixations was the dependent measure, with the AOI defined by each of the six columns (hundreds, tens, ones on the left and right sides of the screen). We found main effects for all variables except for side of screen. There were also several significant interactions. The key findings are highlighted in Table 2. Data characteristics affected fixations: participants fixated

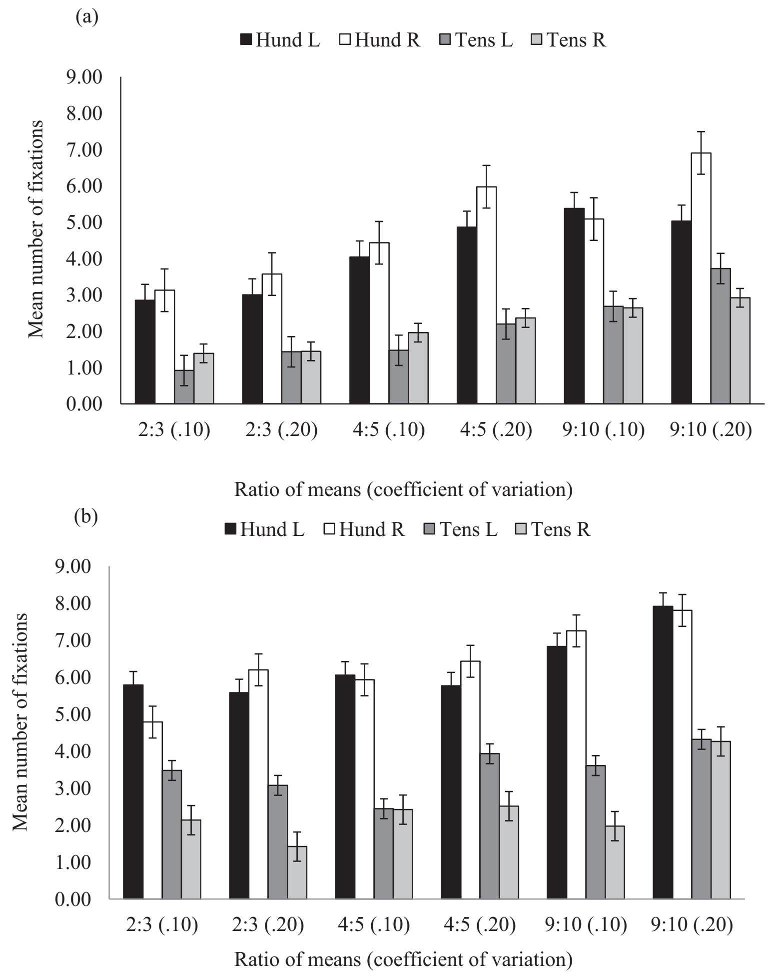

Fig. 2. (a and b) Mean fixation counts on the hundreds and tens column, on the left and right of the screen, by ratio of means and coefficient of variation (in parentheses) with bars representing standard error of the mean. For the sake of simplicity, we do not show fixation count in the ones column, which averaged less than 1 in every single condition. Panel a displays results for sets of four observations. Panel b displays results for sets of eight observations.

more often before making a decision when the means were closer together, and when the data were more variable. Not surprisingly, when there were more data points, there were also more fixations. Finally, the most dramatic effect was that there were an average of 5.5 fixations in the hundreds column, 2.5 fixations in the tens column, and 0.5 fixations in the ones column. Many of the interactions indicate that when low-contrast characteristics are combined, there is an even larger effect; for example, there was a bigger difference in number of fixations based on coefficient of variation when the means were closer together. Further, although there was no main effect of the side of the screen the data

Table 2
ANOVA explanation table

Variables	F	p	Effect Size (Partial $\eta 2$)	Description of Pattern
Main effects				
Mean ratio	78.8	<.001	.88	More fixations with closer ratios
Coefficient of variation	56.1	<.001	.71	More fixations with greater coefficient of variation
Sample size	44.7	<.001	.66	More fixations with 8 pairs than 4 pairs
Left/right column	2.6	.121	.10	No difference
Hundreds/tens/ones	413.9	<.001	.97	More fixations on hundreds than tens or ones; more fixations on tens than ones
Significant 2-way interactions				
Mean ratio × coefficient of variation	19.6	<.001	.64	More fixations based with higher coefficient of variation, but only when the means are closer together
Mean ratio × sample size	13.2	<.001	.55	For smaller sample size, there is a bigger difference in number of fixations based on mean ratio
Mean ratio × left/right	4.1	.030	.27	With lower coefficient of variation, there are more fixations on the left than the right.
Mean ratio × hundreds/tens/ones	27.8	<.001	.85	More fixations on hundreds and tens when the means are closer together than farther apart
Coefficient of variation × left/right	5.1	.034	.18	When coeff. of variation is smaller, more fixations on left than right; more even with greater coefficient of variation
Coefficient of variation × hundreds/tens/ones	26.3	<.001	.71	Fixations on hundreds and tens more affected by coefficient of variation than on ones
Sample size × left/right	11.8	.002	.34	At sample size 8, there are more fixations on the left; at sample size 4, it's even
Sample size × hundreds/ten/ones	36.9	<.001	.77	More fixations in larger set for hundreds and tens columns, but not ones column
Significant 3-way interactions				
Mean ratio × sample size × left/right	8.8	<.001	.64	
Sample size × left/right × hundreds/tens/ones	14.4	<.001	.57	
Coefficient of variation × left/right × hundreds/tens/ones	5.6	.011	.34	

were presented on, this variable did interact with several other characteristics. When the sets were easier to distinguish (higher contrast), there tended to be more fixations on the left side than the right, while when the data were more difficult to distinguish (lower contrast), the fixations were evenly distributed between sides. The first fixation was on the left side 97% of the time, suggesting that there was less need to look back and forth as

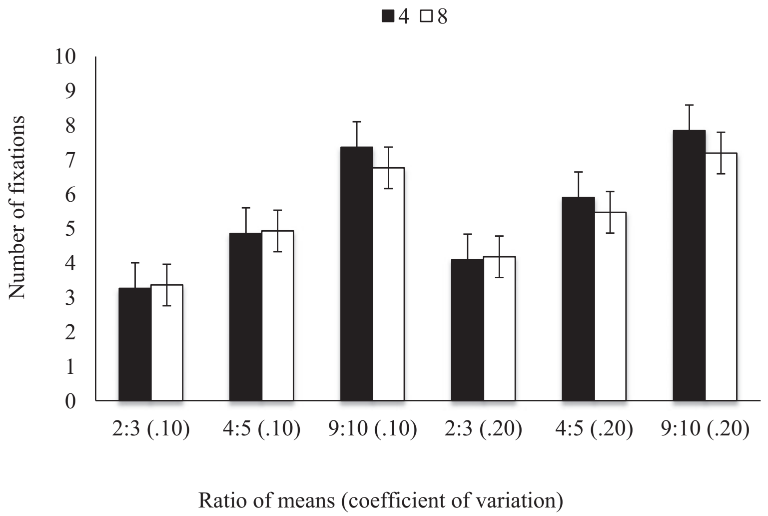

Fig. 3. Number of column switches by ratio of means, coefficient of variation, and set size. *Note.* Error bars represent standard error of mean.

much with sets that are higher contrast. An additional ANOVA was run with fixation duration as the dependent measure, and it yielded similar pattern of results, with longer fixations on the hundreds column as compared to the tens or ones columns, when the means were closer together, and when the coefficient of variation was greater.

3.4. Column switches

Another measure of comparison was the number of times people moved their eyes from one data set to the other. Two research assistants examined video of eye movements and noted how many times fixations switched either from left to right, or right to left. The two coders agreed 92% of the time and discrepancies were resolved through discussions with the first author. We then examined the effect of data characteristics on the frequency of fixation switches. A mean ratio (2:3, 4:5; 9:10)× coefficient of variation (.10 or .20 of mean)× sample size (4, 8) repeated measures ANOVA was conducted with the number of fixation switches as the dependent measure. There were more fixation switches when the means were closer (2:3 ratio: $M = 3.73$, $SD = .51$; 4:5 ratio: $M = 5.29$, $SD = 0.85$; 9:10 ratio: $M = 7.26$, $SD = 1.10$) than when they were farther apart, (F (2, 22) = 164.98, $p < .001$, partial $\eta^2 = .94$). Post hoc paired t tests indicate that the differences were significant between each pair of ratios (all ps < .001). There were also more fixation switches when the coefficient of variation was larger ($M = 5.78$, $SD = 0.78$) than when it was smaller ($M = 5.08$, $SD = 0.77$), (F (1, 23) = 62.64, $p < .001$, partial $\eta^2 = .73$). Finally, there was no main effect of sample size, but at a ratio of 9:10, there were more fixation switches when there were four pairs of data ($M = 7.62$, $SD = 1.63$) than eight ($M = 6.90$, $SD = 1.02$), (F (2, 22) = 3.45, $p = .05$, partial $\eta^2 = .24$; see Fig. 3).

4. Discussion

The results strongly suggest that adult participants were comparing datasets by estimating a summary value that includes information about both means and variance of sets. Accuracy and confidence were higher with high-contrast sets than with low-contrast sets. These results also provided critical information about how participants compared datasets. The number of fixations increased as set contrast decreased. Further, the number of column switches, scans back and forth between data columns, increased as set contrast decreased. These additional fixations were focused on additional set information, commonly the tens column and fixations between individual set values (i.e., whole numbers). It is reasonable to interpret these fixations as seeking additional set details before producing a comparison. When the estimates yield diagnostic summary values and are sufficient to differentiate sets (e.g., higher contrast sets) then there is no need for further action. But when estimates yield non-diagnostic summary values (e.g., lower contrast sets), additional processing beyond the initial estimate occurs. The results of this experiment demonstrate that these additional processes involve attending to new information (e.g., tens values) and making comparisons between whole numbers in each set. Finally, confidence was lower when the number of fixations increased, although both are likely artifacts of the set properties.

Our results are consistent with previous research demonstrating that number comparisons are made on the basis of approximate magnitudes rather than exact values. Our results extend these findings in that differences between sets are detected quickly through the comparison of approximate summary values. Although mean differences appear to be the primary factor in the summary representations (mean differences demonstrated the largest main effect sizes), set variance consistently influenced results indicating that both properties were included in the summary representations.

The results also provide some evidence that multi-digit numbers are decomposed during comparisons. Eye fixations demonstrated that participants attended most frequently to the hundreds column and attention to other place values increased as the contrast between sets decreased (similar to Moeller et al., 2009). Finally there was an increase in the number of column switches as the contrast between sets decreased, consistent with a search for additional information. This finding is similar to results from reading literature in which regressions (i.e., visual scans of previously scanned information; Rayner & Pollatsek, 1989) are an index of comprehension difficulty and suggests an index of the level of comparison difficulty.

4.1. A model of summary representation

We suggest that multiple numbers might be compared in two steps: (a) by creating a summary value for each set via *summary activation areas* along the number line and (b) comparing these summary values. Summary values appear to include information about means and variance. Each individual value creates an activation area, and the overlap between these areas would summarize a set of numbers. The resulting activation areas would implicitly represent means and variance differently. Means would be represented "vertically"—in a perfectly symmetrical set, the mean would be represented as the highest peak in a normal curve (see Fig. 1). Variance would be represented "horizontally" measured as spread of activation across the number line. The difference between summary values would be compared on a number line, similar to the comparisons between single values (Diester & Nieder, 2007).

Our results allow us to evaluate and increase the detail in our model. As outlined above, the process of creating summary values and comparison is initially sequential (Dehaene, 2009); however, as more detail is required (e.g., given low-contrast sets or need for accuracy), additional information may be added to the initial estimates. The eye fixation data demonstrated that first fixations were most commonly directed to the hundreds column of each set. If the resulting scan produces two activation areas that are clearly different (i.e., far apart on the number line), then a comparison can be made without further information. Comparisons between high-contrast sets (e.g., sets with 2:3 ratio of mean difference) were associated with fewer fixations than comparisons between low-contrast sets. If the initial approximate representation is not clearly different, then more information is added to the initial estimate. Subsequent scans are more likely to include additional information not gathered in the initial search, such as looking at the tens places, or returning to numbers previously examined, in order to create a more detailed representation for comparison. In addition to increasing the amount of precision in the approximate representations, additional scanning may also maintain the information already in working memory. This is consistent with evidence from the larger number of fixations on lower contrast sets, more fixations outside of the hundreds column, and more column switch fixations. Comparisons between summary values would produce the same types of distance effects seen in single value comparisons in that larger differences (e.g., higher contrast sets) would be faster and more accurate than smaller differences (e.g., lower contrast sets). Our results also suggested that comparisons that were not sufficiently different triggered a search for additional information.

Our results suggest that there is no single "threshold" at which more information would be needed. Instead, they suggest an inverse, linear relation in which the need for information increases as the contrast between sets decreases. The results suggest the properties of sets that make comparisons relatively easy. High-contrast sets with high mean differences and low variance are relatively easy for adults to compare accurately. Our results also begin to suggest a threshold at which such comparisons become increasingly difficult. When the ratio of mean differences reached 9:10, accuracy decreased and reaction times increased as people sought more information before making comparisons. Future research should further examine data characteristics that influence the threshold of comparison. One suggestion is to examine the influence of outliers on comparisons. A second suggestion is to examine other possible data characteristics such as skew. Future research should also link simple comparisons to more complex cognitive operations such as data interpretation and decision making (see Obrecht et al., 2007; Obrecht, Chapman, & Suárez, 2010).

5. Conclusion

Our results demonstrate that adults compare datasets by rapidly estimating summaries that include statistical properties of the datasets. Comparisons were not based on single values but were influenced by set means and set variance. When sets were highly distinct subjects made an initial scan of the set, usually focusing on the hundreds column of each set, and made rapid and accurate comparisons. As sets became more similar, reasoners increased the number of scans and the type of information sought (e.g., tens column), and confidence and accuracy decreased. Our results extend information about the cognitive architecture for number suggesting that reasoners summarize multi-digit number sets without deliberate calculation, which allows for quick and accurate comparisons when sets are distinct.

Acknowledgments

The authors thank Dave Klein, Bob Khatami, Adeel Raja, and Vince Brown for help with stimulus creation, and John Opfer, Shannon Zentall, and Kristin Weingartner for comments on earlier drafts.

Notes

1. We follow Dehaene (2009) in positing logarithmic number representations with constant variability. For a contrasting view suggesting linear number representations with scalar error variance, see Leslie, Gelman, and Gallistel (2008) and Whalen, Gallistel, and Gelman (1999).
2. We chose to focus on only accurate trials as we were most interested in the effects on behavior and confidence when participants were accurate. However, running similar analyses on the complete dataset including both accurate and inaccurate datasets yields identical patterns of significant main effects and interactions for all of the subsequent analyses reported.

References

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

Buckley, P. B., & Gillman, C. B. (1974). Comparisons of digits and dot patterns. *Journal of Experimental Psychology*, *103*, 1131–1136. doi:10.1037/h0037361.

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

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

Dehaene, S. (2001). Précis of the number sense. *Mind and Language*, *16*, 16–36.

Dehaene, S. (2009). Origins of mathematical intuitions: The case of arithmetic. *The Year in Cognitive Neuroscience 2009: Annual New York Academy of Science*, *1156*, 232–259. doi: 10.1111/j.1749-6632.2009.04469

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 626–641. doi:10.1037/0096-1523.16.3.626.

Diester, I., & Nieder, A. (2007). Semantic associations between signs and numerical categories in the prefrontal cortex. *PloS Biology*, *5*, 2684–2695.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307–314. doi:10.1016/j.tics.2004.05.002.

Grossberg, S., & Repin, D. V. (2003). A neural model of how the brain represents and compares multi-digit numbers: Spatial and categorical processes. *Neural Networks*, *16*, 1107–1140. doi:10.1016/S0893-6080(03)00193-X.

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

Korvorst, M., & Damian, M. F. (2008). The differential influence of decades and units on multidigit number comparison. *The Quarterly Journal of Experimental Psychology*, *61*, 1250–1264. doi:10.1080/17470210701503286.

Leslie, A. M., Gelman, R., & Gallistel, C. R. (2008). The generative basis of natural number concepts. *Trends in Cognitive Sciences*, *12*, 213–218.

Libertus, M. E., & Libertus, K. (2011). Differences in strategies during approximate number comparisons as revealed by eye-gaze measures. Poster presented at the annual meeting of the Association for Psychological Science, Washington, DC.

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.

Moeller, K., Fischer, M. H., Nuerk, H., & Willmes, K. (2009). Sequential or parallel decomposed processing of two-digit numbers? Evidence from eye-tracking. *The Quarterly Journal of Experimental Psychology*, *62*, 323–334. doi:10.1080/17470210801946740.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–1520. doi:10.1038/2151519a0.

Nuerk, H., Kaufmann, L., Zoppoth, S., & Willmes, K. (2004). On the development of the mental number line: More, less, or never holistic with increasing age? *Developmental Psychology*, *40*, 1199–1211. doi:10.1037/0012-1649.40.6.1199.

Nuerk, H., Weger, U., & Willmes, K. (2001). Decade breaks in the mental number line? Putting the tens and units back in different bins. *Cognition*, *82*, B25–B33. 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*, 1147–1152.

Obrecht, N. A., Chapman, G. B., & Suárez, M. T. (2010). Laypeople do use sample variance: The effect of embedding data in a variance-implying story. *Thinking & Reasoning*, *16*, 26–44. doi:10.1080/13546780903416775.

Opfer, J. E., & Siegler, R. S. (2012). Development of quantitative thinking. In K. Holyoak & R. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 585–605). Cambridge, UK: Oxford University Press.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*, 125–157. doi:10.1037/0033-295X.105.1.125.

Sprute, L., & Temple, E. (2011). Representations of fractions: Evidence for accessing the whole magnitude in adults. *Mind, Brain, and Education*, *5*, 42–47. doi:10.1111/j.1751-228X.2011.01109.x.

Trumpower, D. L., & Fellus, O. (2008). Naıve statistics: Intuitive analysis of variance. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 499–503). Austin, TX: Cognitive Science Society.

Van Opstal, F., de Lange, F. P., & Dehaene, S. (2011). Rapid parallel semantic processing of numbers without awareness. *Cognition*, *120*, 136–147. doi:10.1016/j.cognition.2011.03.005.

Vickers, D., Burt, J., Smith, P., & Brown, M. (1985). Experimental paradigms emphasizing state or process limitations: I. Effects of speed–accuracy tradeoffs. *Acta Psychologica*, *59*, 129–161. doi:10.1016/0001-6918(85)90017-4.

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Non-verbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*, 130–137.