STRATEGIC INFERENCE OF MEANS AND VARIANCES: 
AN INVESTIGATION OF ADULT AND CHILD NUMERICAL PREDICTION

A dissertation submitted to the 
Kent State University College 
of Education, Health, and Human Services 
in partial fulfillment of the requirements 
for the degree of Doctor of Philosophy

By

Patrick F. Cravalho

December 2015
A dissertation written by

Patrick F. Cravalho

B.S., Indiana University South Bend, 2007
M.A., San Jose State University, 2010
Ph.D., Kent State University, 2015

Approved by

Bradley J. Morris, Co-director, Doctoral Dissertation Committee

Christopher A. Was, Co-director, Doctoral Dissertation Committee

Clarissa A. Thompson, Member, Doctoral Dissertation Committee

Accepted by

Mary M. Dellmann-Jenkins, Director, School of Lifespan Development & Educational Sciences

Mark A. Kretovics, Dean, College of Education, Health, and Human Services
Everyday we come across the need to make predictions based on sets of numbers. For example, thinking about past prices of gas to estimate how much an upcoming fill up will cost. Prediction is important to study because it represents an essential skill in reasoning with numbers in context, which is a specific component in the Common Core Mathematics Standards. However, little is known about the cognitive mechanisms underlying prediction or how to provide effective instruction. Three experiments investigated the strategies that adults and 4th grade children use to make predictions from sets of four, six, or eight numbers. They were told that each number in each set represented the distance a batter hit a baseball. Then after seeing each set for a limited amount of time, they were asked to predict how far the batter would hit the next baseball. It was hypothesized that predictions would reflect the set means. Adults’ predictions were closer to the set means than children’s predictions, although this difference was not statistically significant. Trying to average the set numbers, which seemed to be the most effective strategy, was the strategy most used by adults. Conversely, children used other strategies more often than averaging and exhibiting a more rudimentary understanding and use of averaging strategies. However, the fact that children employed inexact averaging strategies, without the aid of prior averaging instruction, provides evidence that intuitive number approximation may extend to estimating predictions.
ACKNOWLEDGEMENTS

First, I want to thank my dissertation co-directors, Brad Morris and Chris Was. Brad, thank you for introducing me to the line of number set research, for the time you spend helping me with my various projects, and for your numerical cognition and child development expertise. Chris, thank you for your consultation on my dissertation and related projects, your help recruiting undergraduate participants, and your working memory expertise. I would also like to thank the third member of my dissertation committee, Clarissa Thompson, for her helpful insights into numerical cognition research with children and her introspective feedback on drafts of my dissertation. In addition, I want to thank Lisa Borgerding, the graduate representative of my dissertation committee, for her feedback on late drafts of my dissertation and her expertise on how teaching standards influence learning.

I would like to thank everyone who helped me collect data with local 4th grade classes. First and foremost, I need to thank Annette Kratcoski for putting me in touch with principals and teachers from the Stow-Munroe Falls City Schools. I also need to thank Annette and Tom McNeal for aiding my data collection at the Kent State University Research Center for Educational Technology’s AT&T classroom. Next, I want to thank Principal John Lacoste and 4th grade teachers Linda Laubach and Jerold Dorton from Indian Trail elementary school, as well as Principal David Ulbricht and 4th grader teachers Jamie Delorenzo and Rebecca Miller from Echo Hills elementary school. Last but not least, I want to thank Jenna Wall for her friendship and support with our joint and individual research endeavors including my dissertation.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................ iv

LIST OF FIGURES ............................................................................................................. ix

LIST OF TABLES ................................................................................................................ xi

CHAPTER

I. INTRODUCTION .................................................................................................................. 12
   Overview, Research Questions, and Goals ........................................................................ 12
      Mean Judgments and Number Set Predictions ............................................................. 17
      Strategic Number Set Processing ................................................................................... 18
      Main Hypotheses ............................................................................................................. 19
   Overview of the Methods .................................................................................................. 21
      Participants ...................................................................................................................... 21
      Design .............................................................................................................................. 21
      Setting and apparatus ...................................................................................................... 21
   Prediction Findings ........................................................................................................... 23
      Sets of simultaneously presented numbers ...................................................................... 23
      Sets of sequentially presented numbers .......................................................................... 24
   Strategic Inferences .......................................................................................................... 25
   Educational Implications .................................................................................................. 26

II. REVIEW OF THE LITERATURE ......................................................................................... 28
   Mean Judgments and Number Set Predictions ............................................................... 28
      Early Studies of Mean Judgments .................................................................................... 29
      Extending the Study of Mean Judgments ........................................................................ 31
      Extending the Study of Mean Judgments of Combined Sets ......................................... 33
      Number Set Predictions .................................................................................................. 35
      Mean Judgments, Sampling Strategy, and Working Memory ........................................ 41
      Additional Cognitive Factors in Strategic Numerical Processing .............................. 43
      Strategies for Individual Number Set Processing .......................................................... 46
      Strategies for Number Set Comparisons ......................................................................... 48
   Literature Summary, Research Goals, and Hypotheses .................................................. 49

III. METHODOLOGY .............................................................................................................. 55
   Overview of Experiments .................................................................................................. 55
   Experiment 1 ..................................................................................................................... 55
      Participants ..................................................................................................................... 55
      Design ............................................................................................................................ 57

v
Simultaneously presented numbers: Set size and set variability findings ................................................................. 126
Simultaneously presented numbers: Working memory findings 127
Simultaneously presented numbers: Task order findings ........ 128
Sequentially presented numbers: Descriptive overview .......... 128
Sequentially presented numbers: Set size and set variability findings ................................................................. 130
Sequentially presented numbers: Working memory findings .... 131
Sequentially presented numbers: Task order findings .......... 132
Simultaneous v. sequential sets ........................................ 132
Adult v. Child Predictions ........................................... 133
Simultaneously presented numbers ................................ 133
Sequentially presented numbers ................................ 133
Child Strategy Self-reports ........................................ 133
Summary of typed descriptions .................................... 134
Simultaneously presented numbers: Typed descriptions ........ 135
Sequentially presented numbers: Typed descriptions ........ 135
Summary for Adult Predictions .................................... 139
Summary for Child Predictions .................................... 140
Adult v. Child Predictions ........................................ 141
Child Strategy Summary ........................................ 141

V. DISCUSSION, IMPLICATIONS, AND RECOMMENDATIONS .......... 143
Strategic Inferences ...................................................... 143
Shifts in Strategy Use: Simultaneous Presentation .......... 146
Shifts in Strategy Use: Sequential Presentation .......... 147
Educational Implications .............................................. 149
Limitations and Modifications ........................................ 152
Insufficient Statistical Power ........................................ 153
Inadequacies Common to All Experiments ...................... 154
Inadequacies Common to Experiments 2 and 3 ............... 156
Inadequacies of Experiments 3 ................................... 159
Sequential Version of the Experiment 3 Exemplar Prediction Task .... 161
Summary of Suggested Modifications for Future Research .......... 162

REFERENCES ........................................................................................................ 164
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Example number set prediction task trial (simultaneous presentation). (Slides 1-3) Trial with four numbers; (Slide 4) Open-ended strategy self-report prompt following block with four numbers per trial</td>
<td>59</td>
</tr>
<tr>
<td>2. Example of an ABCD grammatical reasoning task simple statement trial</td>
<td>62</td>
</tr>
<tr>
<td>3. Example of an ABCD grammatical reasoning task challenging statement trial</td>
<td>64</td>
</tr>
<tr>
<td>4. Example of a trial with four numbers from the number set prediction task (sequential presentation)</td>
<td>71</td>
</tr>
<tr>
<td>5. Example of an unmarked 0-1,000 number line from the number-to-position magnitude estimation task</td>
<td>81</td>
</tr>
<tr>
<td>6. Mean Experiment 1 prediction z-scores for sets of simultaneously presented numbers by set size and set variability</td>
<td>87</td>
</tr>
<tr>
<td>7. Mean Experiment 1 response times for sets of simultaneously presented numbers by set size and set variability</td>
<td>89</td>
</tr>
<tr>
<td>8. Mean Experiment 2 prediction z-scores for sets of simultaneously presented numbers by set size and set variability</td>
<td>98</td>
</tr>
<tr>
<td>9. Mean Experiment 2 prediction z-scores for sets of sequentially presented numbers by set size and set variability</td>
<td>100</td>
</tr>
<tr>
<td>10. Mean Experiment 2 response times for sets of simultaneously presented numbers by set size and set variability</td>
<td>103</td>
</tr>
<tr>
<td>11. Mean Experiment 2 response times for sets of sequentially presented numbers by set size and set variability</td>
<td>104</td>
</tr>
<tr>
<td>12. Aggregated adult number line estimations</td>
<td>117</td>
</tr>
<tr>
<td>13. Aggregated child number line estimations</td>
<td>118</td>
</tr>
<tr>
<td>14. Mean Experiment 3 adult prediction z-scores for sets of simultaneously presented numbers by set size and set variability</td>
<td>119</td>
</tr>
</tbody>
</table>
15. Mean Experiment 3 adult prediction z-scores for sets of sequentially presented numbers by set size and set variability ................................................................. 122

16. Mean Experiment 3 child prediction z-scores for sets of simultaneously presented numbers by set size and set variability ................................................................. 126

17. Mean Experiment 3 child prediction z-scores for sets of sequentially presented numbers by set size and set variability ................................................................. 130

18. Mean Experiment 2 and 3 prediction z-scores for sets of simultaneously presented numbers by set size ........................................................................................................ 145

19. Mean Experiment 2 and 3 prediction z-scores for sets of sequentially presented numbers by set size ........................................................................................................ 145
LIST OF TABLES

Table                                                                                           Page
1. Means and Standard Deviations for the Strategy Description Frequency Ratings ...... 93
2. Proportion of Participants Descriptions that described Listed Strategies ................. 111
3. Proportion of Participants Descriptions that described Listed Strategies ................. 138
CHAPTER I
INTRODUCTION

Overview, Research Questions, and Goals

One situation we come across everyday is the need to make judgments or predictions based on categorical sets of numbers (Krueger, Rothbart, & Sriram, 1989; Lovie & Lovie, 1976; Peterson & Beach, 1967). For instance, we may think about past per gallon prices of gas to estimate how much an upcoming fill up will cost. People often make such informal estimations from sets of numbers, which can be thought of as predictions of future number set exemplars. Prediction is an important topic to study because it represents an important skill in reasoning with data, or numbers in context, which is a specific component in the Common Core Standards (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). Even though American mathematics education is currently being aligned with standards that emphasize the representation and interpretation of data throughout grades 1-5 (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010), little is known about the cognitive mechanisms underlying prediction or how to provide effective instruction. There is surprisingly little research on the development of predictions from number sets. Only one prior number set mean judgment study with adolescents (i.e., 14-16 year old students; Lowe, 1971) and a few studies of children’s abilities to compare means from two number sets (Masnick & Morris, 2008;
Morris, Cravalho, Junglen, Was, & Masnick, 2014) have been conducted, but there are no studies of numerical predictions with groups other than adults. The current studies focused on numerical predictions from number sets, which are inferential in nature (Levin, 1976; Levin, Ims, Simpson, & Kim, 1977; Pollard, 1984), but they also relate to mean judgments of number sets, which are descriptive in nature (Peterson & Beach, 1967; Pollard, 1984).

Foundational research by Kahneman and Tversky (1973; “heuristic of representativeness”) posited that humans make non-numerical predictions by selecting the outcome that most represents the set. In the field of non-numerical category representation, multiple authors have likened this prototypical set representation to a statistical characterization of central tendency used to categorize other set members (Bomba & Siqueland, 1983; Medin, Altom, & Murphy, 1984; Nosofsky, Denton, Zaki, Murphy-Knudsen, & Unverzagt, 2012; Redd, 1972). An example would be that people often cite a robin as a prototypical bird because it exhibits all of the common features shared across birds (Medin et al., 1984). In the context of numerical predictions, similar processes may be operating (i.e., abstraction of the prototypical set representation), but will yield prototypes such as the set mean. Prior research shows that adults are very accurate in judging the mean of a number set, but less is known about how close adult number set predictions come to the set means (see Pollard, 1984 for a review) and nothing is known about how children make predictions from number sets. Although the task goal is different, there is evidence to support that mean judgments and predictions are based on analogous mental averaging operations (Krueger et al., 1989; Levin, 1976;
Levin et al., 1977; Malmi & Samson, 1983) that are influenced by one’s inference of number set mean and variance (Peterson & Beach, 1967; Pitz, Leung, Hamilos, & Terpening, 1976; see Pollard, 1984 for a review). Therefore, the current set of studies was guided by the following research question, “Do adults and children infer set means and variances when making predictions of future number set exemplars?”

The inference of set means and variances implies that rather than the proficient “intuitive statisticians” humans were once billed as (Peterson & Beach, 1967, p. 29), we seem to be only “reasonably competent intuitive arithmeticians” (Pollard, 1984, p. 16). Humans are likely strategic in the encoding of set central tendency because of the inherent limitations pointed out by Simon (1959; “bounded rationality”) and others that constrain such processing (Kahneman & Tversky, 1973; Slovic, Kunreuther, & White, 1974; Pitz et al., 1976). Encoding is the preparation of information for storage in memory and later retrieval from memory (Baddeley, Eysenck, & Anderson, 2009). One particular encoding limitation is memory capacity (Estes, 1976; Malmi & Samson, 1983). For example, number set encoding strategies seem to involve averaging over the set members in order to combat the limitations placed on working memory (Cravalho, Morris, Was, & Masnick, 2013; Hyde, 2011). Working memory is often described as the processes and mechanisms involved with the maintenance of task relevant information necessary for performance of cognitive tasks (Miyake & Shah, 1999). Despite multiple prior studies discussing the role of a working memory dependent sampling strategy in number set encoding (Brezis, Bronfman, & Usher, 2015; Malmi & Samson, 1983; Lindskog, Winman, & Juslin, 2013a; 2013b; Lindskog & Winman, 2014; Spencer, 1963),
none of these studies actually measured working memory capacity to investigate how it influenced the numerical predictions of adults or children. Thus, it is not yet clear if the strategies we use to draw inferences about central tendency involve a trade off between our processing costs and the amount of space available to store and process information (Pitz et al., 1976). This leads to the second research question of the current studies, “How does working memory capacity influence adults and children’s predictions of future number set exemplars?”

Another shortcoming of prior number set encoding research is that few prior studies have considered strategies, or plans for achieving goals (Siegler, 1996), for number set processing. Stated differently, approaches for processing number sets might vary by individuals and processing goals (Cravalho et al., 2013). Researchers have speculated that predictions of future number set exemplars are often based on specific strategies meant to optimize one’s inference of categorical averages (Krueger et al., 1989; Levin, 1976; Levin et al., 1977; Malmi & Samson, 1983; Pollard, 1984). For example, consider strategies for drafting players for a fantasy sports team. It is typical for those drafting players to interpret the player’s statistical averages from previous seasons as data from which one might predict which available player they think will perform better during the upcoming season and hence which available player to draft to their team. Besides considering mean values, those drafting players might also consider the season-to-season variance in a player’s statistical averages. For instance, an optimal strategy may be to draft players who are more consistent in their season-to-season performance, knowing they are likely to average close to the same statistics as in prior seasons. A
more risky strategy would be to draft players whose performance vacillates more from season-to-season, but who definitely generate better statistics than more consistent season-to-season performers when they are performing at their best. The implication of the fantasy sports examples is that we make strategic, rather than strictly intuitive (Malmi & Samson, 1983), inferences about a number set’s average via mean and variance information when estimating a prediction from that set. To this end, there is evidence to support that we indeed abstract numerical set central tendency based on the mean and standard deviation of the set (Pitz et al., 1976). Therefore, it follows that one goal of numerical prediction is to extract information about the statistical properties of number sets, such as set mean and set variance, which is achieved by applying appropriate encoding strategies. This leads to the final research question of the current studies, “What strategies do adults and children use to make predictions of future number set exemplars?”

Related to number set encoding strategies, is the fact that few prior studies have considered the educational implications of human’s abilities to process number set means and variances, even though it was long ago suggested that such training should begin with estimating means and variances (Lovie & Lovie, 1976). Strategies for making predictions from sets of numbers should show some consistency across adults and children, because in contrast to other stimuli, symbolic numbers are relatively uniform and unambiguous across people (Krueger et al., 1989; Levin, 1975; 1976). Therefore, the current studies were guided by four goals: (a) to add to the literature of adult numerical predictions, (b) to extend the study of numerical predictions to children, (c) to investigate
how working memory capacity influences the numerical predictions of adults and children, and (d) to examine the estimation strategies used by adults and children to generate numerical predictions. Ascertaining the numerical prediction strategies used by adults is important, but it is just as important to compare children’s numerical prediction strategies to the adult strategies. After the differences in adult and child prediction strategies are better understood, more efficient and effective instruction methods to guide American mathematics elementary teaching standards can be properly developed.

**Mean Judgments and Number Set Predictions**

The earliest studies of adult number set mean judgments found that such judgments were very accurate for sets of 2-digit numbers, although higher set variability had negative effects on confidence in nominal mean judgments (Irwin & Smith, 1956; Irwin, Smith, & Mayfield, 1956) as well as on accuracy of numerical mean judgments (Beach & Swenson, 1966; Lovie & Lovie, 1976; Spencer 1961; 1963). One early study investigated encoding strategies and found that those based on averaging samples of simultaneously presented set numbers were the most effective (Spencer, 1963). The next wave of studies by Anderson (1964; 1968) and Hendrick and Costantini (1970) showed mean judgments from sets of sequentially presented numbers to be biased by how the numbers were sequenced and by response mode. Anderson (1968) also extended the finding that numerical mean judgments from sets of simultaneously presented numbers are accurate and that higher set variability has negative effects on judgment accuracy to 3-digit numbers. Levin’s (1974; 1975) studies showed that adults can accurately judge the combined mean of two sets of simultaneously presented numbers. However,
participants provided socially favorable responses related to each study’s stimuli. Specifically, for Levin’s earlier (1974) study participants underestimated the means for sets of prices, whereas for his later (1975) study they overestimated means for sets of IQ scores. Therefore, Levin’s (1974; 1975) findings suggested that mean judgments from sets of simultaneously presented numbers are, like judgments from sets of sequentially presented numbers, also susceptible to bias.

Fowler (1975) was the first to conduct one of four previous number set prediction studies. He found that adult predictions from simultaneously presented set numbers overestimated the set means when the numbers were labeled as prices, but not when the numbers were unlabeled (Fowler, 1975). Pitz et al. (1976) found that arithmetic means and standard deviations were most indicative of participants representation of the sets of sequentially presented numbers, but that their mean inferences had a more pronounced effect on their nominal predictions. Levin (1976) found that “deviant” set members (i.e., outliers) are more noticeable when the set numbers are presented simultaneously than when presented sequentially. Finally, Levin et al. (1977) found that our encoding depends less on deviant set members when making predictions than when making mean judgments.

**Strategic Number Set Processing**

Studies by Malmi and Samson (1983), Lindskog, Winman, and Juslin (2013a; 2013b), and Lindskog and Winman (2014) provide further support that adults make accurate numerical mean judgments and that our judgments are more indicative of arithmetic means than modes or medians. In addition, these studies added support that
number set judgment strategies are based on averaging set samples, whether the numbers are presented simultaneously (Spencer, 1963) or sequentially (Lindskog et al.; Lindskog & Winman, 2014; Malmi & Samson, 1983). Multiple studies support Hyde’s (2011) suggestion that mean judgment strategies are employed based on one’s working memory capacity (Brezis et al., 2015; Cravalho et al., 2013; Lindskog et al.; Lindskog & Winman, 2014; Malmi & Samson, 1983).

In terms of number set mean comparisons, children and adults who aggregated across set members and then compared the average mean and variance values produced accurate judgments of which sets had the higher means (Masnick & Morris, 2008; Morris et al., 2014; Morris & Masnick, 2015). However, children are less accurate and use more strategies than adults, some of which are less efficient (Masnick & Morris, 2008; Morris et al., 2014). This is most likely due to children having less knowledge and experience with number properties and comparisons than adults (Morris et al., 2014).

**Main Hypotheses**

Various hypotheses follow from the previous mean judgment and prediction studies. First, because previous research shows that both mean judgments and predictions by adults are closer to the arithmetic means of number sets than any other measure of central tendency (Lindskog et al., 2013a; 2013b; Lindskog & Winman, 2014; Malmi & Samson, 1983; Pitz et al., 1976), it was hypothesized that the specific measure of central tendency that adult and child number set predictions would reflect is the set mean. Second, because previous research shows that high set variability led mean judgments and predictions by adults to be further from the set means than those from low
variability sets (Anderson, 1968; Beach & Swenson, 1966; Levin, 1976; Levin et al., 1977; Lovie & Lovie, 1976; Spencer 1961; 1963), it was hypothesized that adult and child predictions from sets with high set variability would be further from the set means than those from sets with low set variability. Third, it was hypothesized that children would infer set means and variances, just as adults did when making predictions (Pitz et al., 1976) from number sets. However, it was also hypothesized that children’s predictions would produce values that were further from the set means than adult predictions because they have less experience with numerical encoding than adults (Masnick & Morris, 2008; Morris et al., 2014; Morris & Masnick, 2015). Forth, it was predicted that adults and children with larger working memory capacity would be able to maintain a greater number of items in working memory and more precisely estimate the means of the largest sets included in the current studies (i.e., eight numbers). Finally, based on previous research showing that mean judgments, comparing set means, and predictions share the same task goal of averaging set members (Levin, 1976; Levin et al., 1977; Peterson & Beach, 1967; Pitz et al., 1976; Pollard, 1984), it was hypothesized that predominately averaging strategies would be reported for making number set predictions, similar to reports for mean judgments (Spencer, 1963) and set comparisons (Masnick & Morris, 2008; Morris et al., 2014; Morris & Masnick, 2015). It was also hypothesized that children’s prediction strategies would be less efficient than adult’s. Unlike adults, it was thought that children would use a wider range of strategies without pruning the ineffective tactics as they gain experience with the task, similar to the differences seen
with adult and child number set comparison strategies (Masnick & Morris, 2008; Morris et al., 2014; Morris & Masnick, 2015).

**Overview of the Methods**

In order to test the main hypotheses, participants were first presented with a backstory regarding sets of 3-digit numbers that they were to encode for their predictions. They were told that each number in each set represented the distance in feet a batter hit a baseball during a home run competition. Then after seeing each set for a limited amount of time, they were asked to predict how far the batter would hit the next baseball. This number set exemplar prediction task was designed to induce averaging strategies that may arise in normal (non-laboratory) settings (Malmi & Samson, 1983).

**Participants.** Experiments 1, 2, and 3 included adult participants who were undergraduate students recruited from the Kent State University Educational Psychology course pool of participants. Experiment 3 also included child participants who were 4th grade students recruited from local schools with assistance from the Research Center for Educational Technology’s AT&T Classroom (on the campus of Kent State University). Kent State University intuitional review board (IRB) approval was obtained prior to the recruitment of participants. All standards for ethical treatment of participants set forth by the American Psychological Association (APA), including obtaining informed consent and maintaining confidentiality, were followed at all times.

**Design.** Experiment 1 used a 3 (number set size) x 2 (number set variability) within-subjects design. The experimenter programmed a number set exemplar prediction task, with sets of simultaneously presented numbers, which was used to measure
participant exemplar predictions, response latency, and strategy self-reports (frequency rating and open-ended). The ABCD grammatical reasoning task (cf. Was & Woltz, 2007) was used to measure verbal working memory capacity.

Experiments 2 and 3 used a 3 (set size) x 2 (set variability) x 2 (set presentation) within-subjects designs. In addition to a modified version of the simultaneous presentation (i.e., all the set numbers presented at once) version of the number set exemplar prediction task used in Experiment 1, a sequential presentation (i.e., one set number presented at a time) version of the exemplar prediction task was programmed. Each Experiment 2 task measured exemplar predictions, response latency, and open-ended strategy self-reports. Each Experiment 3 task measured exemplar predictions and eye fixations. For Experiment 2, the ABCD grammatical reasoning task (cf. Was & Woltz, 2007) was again used to measure verbal working memory capacity. For Experiment 3, the child participants were asked to verbally report on their number set encoding strategies and to complete the short version of the automated working memory assessment (AWMA), which measured their verbal working memory capacity. Also for Experiment 3, each adult and child completed a number-to-position magnitude estimation task measuring his or her mental representation of numbers in the 0-1,000 range (i.e., the range of numbers used to create the experimental sets).

**Setting and apparatus.** Experiments 1 and 2, and the adult phase of Experiment 3 took place in the Educational Psychology Laboratory on the campus of Kent State University. This laboratory contained multiple computer stations. All experimental tasks
were loaded onto the stations and participants completed all experimental tasks individually.

The child phase of Experiment 3 took place in multiple locations, either at the AT&T classroom or in the school from which the children were recruited. In the space provided, all necessary equipment to complete all experimental tasks were set up for data collection. All Experiment 3 adult and child participants also completed number-to-position magnitude estimations via a paper and pencil number line task.

**Prediction Findings**

The findings support that both adults and children, regardless of whether the set numbers were presented simultaneously or sequentially, inferred set means and variances when estimating set exemplar predictions.

**Sets of simultaneously presented numbers.** Although not a statistically significant difference, children’s predictions were further from the set means than adult’s predictions. Set size did have a statistically significant influence on how close adult and child predictions were to the set means. The child and adult set size patterns were mostly identical, with predictions becoming further from the set means as set size increased from four numbers to six numbers, and then predictions becoming closer to the set means as set size increased from six numbers to eight numbers. Set variability did not have a statistically significant influence on how close adult or child predictions were to the set means. Finally, a very surprising finding was that having higher working memory capacity made no difference in how close adult or child predictions were to the set means.
Adult predictions were closest to the set means for sets of eight numbers, coming within 3 digits, on average, of the set means, and furthest from the sets means for sets of six numbers, coming within 12 digits, on average, of the set means. Adult predictions for sets of four numbers were, on average, within 6 digits of the set means.

Child predictions were also closest to the set means for sets of eight numbers, coming within 18 digits, on average, of the set means, and also furthest from the sets means for sets of six numbers, coming within 38 digits, on average, of the set means. Child predictions for sets of four numbers were, on average, within 21 digits of the set means.

**Sets of sequentially presented numbers.** Although not a statistically significant difference, children’s predictions were further from the set means than adult’s predictions for sets of six and eight numbers, but were closer to the sets means than adult’s predictions for sets of four numbers. Set size again had a statistically significant influence on how close adult and child predictions were to the set means. The child and adult set size patterns were identical for sets of four and six numbers, with predictions becoming closer to the set means as set size increased from four numbers to six numbers. However, as set size increased from six numbers to eight numbers, adult predictions became slightly closer to the sets means, whereas children’s predictions became much further from the set means. Set variability again did not have a statistically significant influence on how close adult or child predictions were to the set means. Finally, a very surprising finding was that having higher working memory capacity made no difference in how close adult or child predictions were to the set means.
Adult predictions were closest to the set means for sets of eight numbers, coming within 3 digits, on average, of the set means, and furthest from the sets means for sets of four numbers, coming with 15 digits, on average, of the set means. Adult predictions for sets of six numbers were, on average, within 4 digits of the set means.

Child predictions were also closest to the set means for sets of six numbers, coming within 9 digits, on average, of the set means, and furthest from the sets means for sets of eight numbers, coming within 45 digits, on average, of the set means. Child predictions for sets of four numbers were, on average, within 11 digits of the set means.

**Strategic Inferences**

Children exhibited a more rudimentary understanding and inefficient use of averaging strategies than adults. Children used inexact phrasing to imply they were trying to average the numbers (e.g., they described trying to find a number “in between” or “in the middle of” the set numbers) rather than saying they “tried to average” the set numbers as many adults described. Trying to average the set numbers seemed to be the most effective strategy and adults utilized it most often for both simultaneously and sequentially presented numbers. Conversely, children utilized three other strategies more often than trying to average the numbers for both simultaneously and sequentially presented numbers, half of which were strategies adults did not use at all. Also unlike the adults, children utilized strategies that were either not related to inferring the means or variances of the sets (e.g., looking at only the last few numbers in the set, adding numbers) or that did not seem to be related to the task goal of making a prediction about
the set (e.g., subtracting or counting numbers). Adults did not use any strategies that were unrelated to the prediction task goal.

**Educational Implications**

The fact that the children exhibited a more rudimentary understanding and use of averaging strategies than adults is not very surprising, as 4th grade teachers in Ohio are not required to teach their students about averaging sets of numbers (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). However, the fact that children employed inexact averaging strategies, without the aid of prior averaging instruction, supports that intuitive number approximation is not limited to counting, addition, subtraction, and magnitude comparisons (Dehaene, 2009; Feigenson, Dehaene, & Spelke, 2004; Gilmore, McCarthy, & Spelke, 2007) and may extend to estimating predictions. Prior research supports that infants have inherent systems in place for averaging set members (Bomba & Siqueland, 1983) and for numerical estimation (Feigenson et al., 2004), and the current studies support that the children use these intuitions to formulate their numerical predictions. However, the current studies also support that children are less adept at number set variance processing than adults. Being that the current standards for grades 1-5 mathematics education emphasize representation and interpretation of number sets (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010), a few implications can be highlighted from the current studies to be applied to 4th grade classrooms. One, if we begin teaching children the more efficient estimation strategies employed by adults, their averaging intuitions should enable them to quickly learn and employ those strategies.
Two, because children are less efficient in their processing of number set variance, as we teach children number set exemplar estimation strategies, a focus should be on the most effective ways to infer set variance.
CHAPTER II

REVIEW OF THE LITERATURE

Mean Judgments and Number Set Predictions

The current experiments focused on numerical predictions, which are inferential in nature (Levin, 1976; Levin et al., 1977; Pollard, 1984), but they also relate to mean judgments, which are descriptive in nature (Peterson & Beach, 1967; Pollard, 1984). To foreshadow, the research discussed below shows that adults are accurate in judging the mean of a number set, but less is known about how close number set predictions come to the set means (see Pollard, 1984 for a review). Although the task goal is different, there is evidence to support that mean judgments and predictions are based on analogous mental averaging operations (Levin, 1976; Levin et al., 1977) based on mean and variance inferences (Peterson & Beach, 1967; Pitz et al., 1976; Pollard, 1984). In addition, more recent studies on number set processing strategies indicate that averaging over the set members functions in part to combat the limitations placed on working memory (Brezis et al., 2015; Cravalho et al., 2013; Hyde, 2011). Finally, it is also important to note that, based on the distinction outlined by Kahneman and Tversky (1973), the existing research on number sets includes nominal (i.e., category-based) and numerical (i.e., number-based) judgments and predictions. In the review below, distinctions will be made as to whether a study used nominal (i.e., participant chose between ranges of numbers) or numerical (i.e., participant responded with an exact number) response types.
**Early Studies of Mean Judgments**

Irwin et al. (1956) reported the results of multiple experiments in which college students were asked to make estimations from sets of sequentially presented positive and negative numbers. Specifically, participants were shown numbers (80% 1-digit numbers, 20% 2-digit numbers) out of a set of 500 until they felt ready to judge if the mean value of the entire set population was less than or greater than zero. These were nominal mean value judgments, as participants were instructed to either say “plus” if they thought the value was positive or say “minus” if they thought the value was negative. They were also asked to rate how confident they were in each judgment. Each set number was presented for between 5 s and 10 s and each set differed in mean and variance. Incorrect “plus/minus” judgments were inversely related to confidence. Overall, participants grew more confident in their mean value judgments as set size increased. However, participant confidence was also related to set variance, as they were less confident in judgments of high variance sets, which was reflected by a significantly higher amount of numbers being viewed before judgments of high variance sets.

Over three studies, Spencer (1961; 1963) investigated numerical mean judgments, at times comparing his results to Irwin et al.’s (1956) investigation of nominal mean judgments. Participants were shown sets of simultaneously presented numbers, for 1 s or less per set number, with the sets containing either high or low variance. Spencer (1961; 1963) found that college students were accurate in judging the mean of sets of 10, 15, or 20 2-digit numbers. However, he also found that error in mean judgments increased along with set size and set variability, results that were later replicated by Beach and
Swenson (1966). The finding from multiple studies of mean judgments of sets of simultaneously presented 2-digit numbers that higher set variability lead to less accurate mean judgments (Beach & Swenson, 1966; Spencer, 1961; 1963) seems to match the finding that participants were less confident in mean judgments of sets of sequentially presented 1- and 2-digit numbers with higher set variability (Irwin & Smith, 1956; Irwin et al.). In addition, a study of sets of sequentially presented numbers with similar parameters to the sets of simultaneously presented number studies (Beach & Swenson, 1966; Spencer, 1961; 1963) indicated that participants mean judgments were close to the actual means, with error increasing along with set variance (Lovie & Lovie, 1976).

After his initial studies, Spencer (1961) hypothesized that Irwin et al.’s (1956) participants’ confidence grew with set size because they were given more contextual information for their judgments (i.e., they knew the exact mean value was close to zero). Therefore, Spencer claimed it was easier to match data to the given context than generate a mean value with no context as his participants had to do. As part of a later study, Spencer (1963) asked participants to report on their averaging strategy to assess how they generated a mean value without context. Three strategy categories were conveyed by the reports: (a) averaging the highest and lowest values, (b) averaging samples of the sets, and (c) averaging based on the 10s column of the numbers. The sampling strategies lead to estimations that were closest to the arithmetic means of the sets. Also, similar to the studies by Irwin et al., participants cited high set variance for their lack of confidence in their estimations.
Overall, set mean judgments were accurate, although higher set variability had negative effects on confidence in nominal mean judgments (Irwin & Smith, 1956; Irwin et al., 1956) as well as on accuracy of numerical mean judgments (Beach & Swenson, 1966; Lovie & Lovie, 1976; Spencer 1961; 1963). In addition, strategies based on averaging samples of simultaneously presented set numbers were the most effective (Spencer, 1963). To summarize, mean judgment studies of sets of simultaneously and sequentially presented numbers produced similar results, but strategies were only investigated for the former.

**Extending the Study of Mean Judgments**

Anderson (1964) conducted the earliest versions of sequentially presented 2-digit number experiments with numerical mean judgments. In one experiment he had college students study sets of eight sequentially presented 2-digit numbers (7 s per number), making an estimation of the set average after each set number was presented. Participant estimations were accurate, but there were clear recency effects based on how the set numbers were sequenced. In a second study, Anderson (1964) again had participants make an estimation of the set average after each of eight set numbers was presented, but he added more sets and a 15 s interval between each set presentation during which participants rated short cartoons. For this study, participant estimations exhibited a serial position effect. Specifically, the set average estimations reported after the sixth number in the set was presented depended on the first five numbers that were presented.

Anderson (1968) also performed studies with sets of simultaneously presented 3-digit numbers. College students were shown two sets of 3-digit numbers and asked to
estimate the combined average of the two sets. The sets ranged from one number to six numbers and were shown for 5 s at a time, which was similar to the time intervals used by Spencer (1961; 1963) and Beach and Swenson (1966) in their studies of sets of simultaneously presented 2-digit numbers. Like those 2-digit studies, Anderson (1968) found that estimation error rate increased with set variance, though participants generally overestimated their mean judgments.

In direct response to Anderson’s (1964) earlier finding of a recency effect for estimations from sets of sequentially presented numbers and a continuous mode of averaging (i.e., estimating the set average after each set number was presented), Hendrick and Costantini (1970) sought to induce a primacy effect with sets of sequentially presented numbers via a final mode of averaging (i.e., estimating the set average after all the set numbers were presented) as used in the previously discussed mean judgment studies (Beach & Swenson, 1966; Spencer 1961; 1963). Across two experiments, college students were shown sets of six numbers, consisting of a series of three 1- and 2-digit numbers and a series of three 3-digit numbers. Each set was presented twice, once with the series of 1- and 2-digit numbers presented first and once with the series of 3-digit numbers presented first. Participants saw each set number for 2 s and after all six set numbers were presented, they were asked to estimate the set average. The results of each experiment showed primacy effects in the participant mean estimates. Specifically, mean estimates for sets where the first three numbers were 1- and 2-digit numbers were systematically lower than the actual set means and mean estimates for sets where the first three numbers were 3-digit numbers were systematically higher than the actual set means.
Overall, the studies by Anderson (1964) and Hendrick and Costantini (1970) showed mean judgments from sets of sequentially presented numbers to be biased by how the numbers are sequenced and by response mode. In addition, Anderson (1968) extended the findings that numerical mean judgments from sets of simultaneously presented numbers are accurate and that higher set variability has negative effects on judgment accuracy to 3-digit numbers. To summarize, mean judgments of 2- and 3-digit numbers produced similar results, but judgments from sets of sequentially presented numbers are more susceptible to bias.

**Extending the Study of Mean Judgments of Combined Sets**

Similar to Anderson (1968), Levin (1974) asked college students to study two sets of simultaneously presented numbers and then judge the combined mean of the sets. In one experiment, participants saw two sets of ten 2-digit numbers for 10 s, being told that each set represented a random sample of IQ scores from an elementary school. The participants were expected to judge the mean IQ of the school based on their mean estimates for the two samples. In a second experiment, participants saw two sets of differing sizes (e.g., 10 and 15 numbers) and were told they belonged to two classes of elementary students. Then they were asked to judge what the mean IQ would be if the two classes were combined. The results for these two experiments were similar as mean judgments were close to the actual population means. However, participants systematically overestimated the set means, which corroborated Anderson’s (1968) results. In the second experiment, when the samples were unequal in size, participant judgments relied more heavily on the larger of the samples. Pollard (1984) pointed out
that Levin’s (1974) participants may have been weighing the two samples as a proportion of the combined population and therefore it would be optimal to favor the larger sample.

Levin (1975) conducted another series of experiments asking participants to judge means based on two samples. Participants saw two sets of five simultaneously presented 2-digit numbers for 10 s, being told that each set represented prices for a certain type of grocery item. The participants were expected to judge the mean price of the two samples. Manipulations across the studies included adding labels to each sample, varying the distance between sample means, and varying the size of the samples (four, eight, 16 or 32 numbers). Similar to Levin’s (1974) previous experiments, mean judgments were close to the actual means, regardless of the manipulation. Mean estimation error was found to increase with variance, which was consistent with research on mean judgments of an individual set of simultaneously presented 2-digit numbers (Anderson, 1968; Beach & Swenson, 1966; Spencer, 1961; 1963). A difference between Levin’s later (1975) and earlier (1974) studies was that participants typically underestimated the sample means rather than overestimate them. Levin (1975) explained that this difference most likely was due to participants favoring certain responses due to personal disposition. For instance, it is socially favorable to overestimate one’s IQ (Levin, 1974) and underestimate price increases (Levin, 1975). This explanation of the conflicting bias seen in Levin’s (1974; 1975) studies reflects reports of such bias from multiple studies of mean, variability, and proportion estimation (see Pollard, 1984 for a review). Finally, when the samples were unequal in size, participant judgments relied more heavily on the larger of the samples, which replicated Levin (1974), but the manipulation of sample
mean also played a role. When the larger sample also had the higher mean, estimations were higher than the actual means, but when the larger sample had the lower mean, estimations were lower than the actual means (Levin, 1975).

Overall, Levin’s (1974; 1975) studies showed that adults can accurately judge the combined mean of two sets of simultaneously presented numbers. However, due to participants providing socially favorable responses related to each study’s stimuli, participants typically underestimated the set means in Levin’s earlier (1974) study and they typically overestimated the set means in his later (1975) study. To summarize, judgments from sets of simultaneously presented numbers are also susceptible to bias.

**Number Set Predictions**

Fowler (1975) showed 36 sets of 12 simultaneously presented 2-digit numbers, varying in mean and range, to college students. The students were randomly split into two groups. One group (price label group) was told that each set represented the past prices of an item. It was then their task to estimate the 24th change in price rather than the 13th change in price so that they would be forced to speculate on future price fluctuations. The other group (no label group) made predictions in the same fashion as the first group, but they were not told that the numbers in each set represented the prices of an item, leaving the numbers unlabeled. No study time limits were specified, but the participants were told to avoid lengthy deliberation for their estimations. The results showed that the no label group’s predictions were, on average, within .1 of the set means, whereas the price label group’s predictions were, on average, within .5 of the set means. The price label group’s predictions overestimated the set means, which the authors
thought reflected a general public bias to assume that prices of most items will increase over time.

Pitz et al. (1976) conducted a study with similar parameters to the mean judgment studies discussed above, but asked participants to make nominal predictions rather than numerical mean estimations. College students were shown sets of nine sequentially presented 3-digit numbers. Each set number was displayed for 3 s and (after the whole set was presented) participants were given 10 s to encode the numbers in any way they saw fit. After the encoding period, participants predicted which of three numerical ranges would contain the next value. The results showed that (given normally distributed samples) the arithmetic means and standard deviations of the sets were the best fit for the central tendency and variability of the sets. However, mean inferences had a more pronounced effect on predictions than variability inferences.

In addition to his descriptive statistical judgment (i.e., mean judgment) studies described earlier, Levin (1976) also investigated inferential statistical judgments through predictions from sets with a “deviant” number, or a number between 13-37 digits of the set mean, whereas the other set numbers where within 10 digits of the set mean. Levin (1975; 1976) assumed that adults assign a psychological weight, or representation of importance, to each number in the sample. Levin (1976) described three ways adults may weigh deviant set information: (a) assign it more weight than the other set members (“overweight” it), (b) weight it equally to the other set members, or (c) assign it less weight than the other set members (“discount” it). To investigate these possibilities, Levin (1976) conducted two similar experiments with college students.
For experiment one, participants saw 75 sets of nine simultaneously presented 3-digit numbers for 12 s per set. The participants were split into four groups. The descriptive group was told that the numbers represented the percentage price increases for a single grocery store item and they were to estimate the mean percentage price increase for each set. The discounting group was told that the numbers represented a random sample of percentage price increases for items from the same grocery store and were to estimate the mean percentage price increase for the store. The discounting group was also explicitly told to disregard any deviant set members. There were two inference groups, both of which were told the same backstory for the set numbers as the discounting group and asked to estimate the mean percentage price increase for the store. However, instead of being instructed to disregard deviant numbers, one inference group was given no further instruction and the other was reminded that not all set members equally represent the store’s price increases. The results provided two clear findings. One, when asked to make a mean judgment (descriptive group only), participants did not discount deviant set members. Two, participants in each of the three groups asked to make a prediction did discount deviant set members, even those who received no briefing related to such outliers. These findings mirrored those of Anderson (1968), who performed studies with simultaneously presented 3-digit number sets and found that those making inferences discounted outliers in a similar fashion to participants in Levin’s (1976) study.

Levin’s (1976) second experiment was exactly like his first experiment except for the following changes. The sets of numbers were presented sequentially instead of
simultaneously and there was only one inference group. The inference group was told
the same backstory for the set numbers as the discounting group and asked to estimate the
mean percentage price increase for the store and given no further instruction. In addition,
each group was split into two subgroups. One subgroup was called the running average
(RA) group. These participants provided a response after seeing each set member (7 s
viewing time per number), similar to Hendrick and Costantini’s (1970) continuous mode
of averaging. The other subgroup was called the end-only (EO) group. These
participants provided a response only after seeing the last set member (3 s viewing time
per number), similar to Hendrick and Costantini’s (1970) final mode of averaging. Each
subgroup was given 12 s to provide a response. The results indicated that participants in
the RA subgroups only discounted when prompted to do so, whereas participants in the
EO inference subgroup showed a slight amount of unprompted discounting of deviant set
members. Hence, deviant set members seemed to be more noticeable when numbers are
presented simultaneously than sequentially (Levin, 1976), contradicting the previous
assertion by Peterson and Beach (1967). In addition, participant responses in the RA
subgroup of the descriptive group reflected a strong recency effect, which is consistent
with the results of Anderson’s (1964) mean judgment task that also required participants
to provide a response after each sequentially presented set number. However, participant
responses in the EO subgroup of the descriptive group did not reflect any recency effect,
which contradicted Hendrick and Costantini’s (1970) results, although the latter’s
procedure included much longer exposure and response times.
Levin et al. (1977) extended Levin’s (1976) investigation of descriptive and inferential statistical judgments through nominal evaluations and predictions from sets with a “deviant” number. Participants saw 27 sets of eight sequentially presented 2-digit numbers, seeing each number for 3 s. The sets were said to represent the percentage correct for differing student’s test scores, with some sets containing a deviant score and some not containing a deviant score. Across all sets, the first half of scores was, on average, higher than the second half. Those in the evaluation group were split into two subgroups. The rating subgroup was told to consider all set scores to be equally important, whereas the rating plus trend subgroup was told to consider whether or not the scores reflected improvement over time. Those in the prediction group were also split into two subgroups. The prediction and the prediction plus discounting subgroups were told to predict each student’s comprehensive exam performance. However, only the prediction plus discounting subgroup were told that unusually high or low-test scores are not the best indicator of future performance. Participants had 7 s to either evaluate or predict using a 20-point scale labeled from “very bad” to “very good”. For the sets without deviant scores, larger recency effects were seen with the prediction groups than the rating subgroup, but not the rating plus trend subgroup. For the sets with deviant scores, the two predictions subgroups and the rating plus trend subgroup showed large recency effects, but the rating subgroup showed only a small recency effect. Levin and colleagues (1977) asserted that the recency effects were due to the embedded trends in the sets. In the case of the prediction subgroups, this supports that participants were actively processing the most recent prior information to make their predictions (Jones,
1971). Also, for the sets with deviant scores, the prediction subgroups reliably discounted any deviant scores, whereas the rating plus trend subgroup only slightly discounted deviant scores and the rating subgroup over-weighted deviant scores. These discounting results reflect that deviant set members were given less weight when making predictions than when making evaluations (e.g., mean judgments). This supports the conclusion that task goal determines how adults encode and process deviant set information (Levin et al., 1977).

Of the four previous number set prediction studies, only Fowler’s (1975) study prompted numerical predictions from sets unbiased by an outlier. Fowler (1975) found that adult predictions from simultaneously presented set numbers overestimated the set means when the numbers were labeled as prices, but not when the numbers were unlabeled. Pitz et al. (1976) found that the arithmetic means and standard deviations of sequentially presented, set numbers were closest to the participants representation of the sets, but that their mean inferences had a more pronounced effect on their nominal predictions. Levin (1976) found that deviant set members are more noticeable when the set numbers are presented simultaneously than when presented sequentially. Levin et al. (1977) found that adults weigh deviant set members less when making predictions than when making mean judgments. To summarize, nominal and numerical predictions reflect set means and variance and outliers bias predictions from sets of simultaneously presented numbers more than from sets of sequentially presented numbers.
Mean Judgments, Sampling Strategy, and Working Memory

Malmi and Samson (1983) conducted three studies during which they presented college students with two randomly interleaved sets of 3-digit numbers (50 numbers per set). Each number was shown for either .5 s or 1.5 s and was presented with one of the two group labels. The participants were told that each number represented a student’s SAT score and that they were to estimate the average SAT score for each group of students. Similar to the prior studies using 2-digit numbers (Anderson, 1964; Beach & Swenson, 1966; Lovie & Lovie, 1976; Spencer, 1961; 1963), estimations were accurate. The analyses for one of the studies showed that the estimations were much closer to the set means than the set modes or medians and that lower between set variance and shorter presentation times caused systematic underestimation of the set means. Another of these studies varied whether participants made their estimations after viewing all or only a segment of the set numbers. The results of this version of the task showed that estimates made based on seeing segments of the set numbers, rather than those made from viewing all the set numbers, were closer to the set means.

Malmi and Samson (1983) purported that these findings support a “fulcrum” hypothesis about how mean estimations are processed. This hypothesis states that humans only store and represent a sample of the set numbers during processing in memory and when someone is asked to estimate the average of the set they assess the “balance point” (i.e., mean) of the sample by mentally weighting the values (Malmi & Samson, 1983). Malmi and Samson’s (1983) hypothesis is consistent with Cowan’s (1995) theory of working memory. Cowan (1995) suggested that working memory is
limited (i.e., roughly four items) and that current representations are only a sample of representations from long-term memory. To frame Malmi and Samson’s (1983) hypothesis using Cowan’s (1995) theory, when asked to make a judgment based on a previously viewed set of numbers, only a sample of the set numbers from our long-term memory will be available for processing due to limited working memory capacity. Malmi and Samson’s (1983) results supported that sampling strategies lead to more accurate set mean estimates from sets of sequentially presented numbers. The sampling strategy hypothesis is also supported by Spencer’s (1963) finding that sampling strategies led to estimations that were closest to the means of sets of simultaneously presented numbers.

More recently, the proposed naïve sampling model (NSM) for data point estimation (Juslin, Winman, & Hansson, 2007; Lindskog et al., 2013a; 2013b; Lindskog & Winman, 2014) agreed with the main tenets of the “fulcrum” hypothesis, that adults store only a sample of numbers from a set to make number set property value judgments (Malmi & Samson, 1983). The NSM is also consistent with predictions from Cowan’s (1995) theory of working memory. According to the NSM (Lindskog et al., 2013a; 2013b), adults store experienced numerical data points (exemplars) in long-term memory. When adults are asked to make a judgment about experienced numbers, processing is constrained by working memory capacity (4 ± 2 observations; Cowan, 2001). Consequently, adults are forced to use a sample to represent the chosen property (e.g., mean, variance, distribution shape, etc.) of the population they are thinking about. The NSM predicts that set mean judgments will be more accurate than set variance judgments
because the mean is an unbiased property under random sample conditions (Lindskog & Winman, 2014). Multiple studies by Lindskog, Winman, and Juslin (2013a; 2013b) provide evidence that adults are accurate in judging the means of sets of sequentially presented 3-digit numbers, which replicates the prior studies discussed above (Anderson, 1964; Beach & Swenson, 1966; Lovie & Lovie, 1976; Malmi & Samson, 1983; Spencer, 1961; 1963).

Overall, studies by Malmi and Samson (1983), Lindskog, Winman, and Juslin (2013a; 2013b), and Lindskog and Winman (2014) provide further evidence that adults make accurate numerical mean judgments and that our judgments are more indicative of the set means than the set modes or medians. In addition, Malmi and Samson (1983), Lindskog et al., and Lindskog and Winman (2014) provided evidence that adults use a sampling strategy to make number set judgments from sets of sequentially presented numbers and theorized that working memory capacity limitations motivate the sampling strategy. In summary, more recent research added support that number set judgment strategies are based on averaging set samples, whether the numbers are presented simultaneously (Spencer, 1963) or sequentially (Lindskog et al.; Lindskog & Winman, 2014; Malmi & Samson, 1983).

**Additional Cognitive Factors in Strategic Numerical Processing**

A recent set of studies asked participants to average sets of sequentially presented 2-digit numbers, with each set number being presented for 500ms or less (Brezis et al., 2015). These studies found that adults were more accurate in estimating the set average from sets of four numbers than from sets of eight numbers, but were more accurate in
estimating averages for sets of 16 numbers than for sets of eight numbers. Brezis et al. (2015) claimed this pattern was a result of a shift in processing strategy as set size increased. Due to working memory limitations, “analytic” processing (i.e., individuating set members) decreased with set size, as “intuitive” processing (i.e., averaging across set members) increased with set size (Brezis et al., 2015). The increased use of an averaging strategy for mean judgments of increasing set size reflects findings that adults (Homa & Vosburgh, 1976), and even 3-4 month old infants (Bomba & Siqueland, 1983), rely more on set prototype knowledge for categorization tasks as the number of exemplars increases. Following Dehaene and colleagues, Brezis et al. (2015) also posited that the non-symbolic approximate number system (ANS) underlies such trends in symbolic numerical estimation (Dehaene, 2001; Lerner, Dehaene, Spelke, & Cohen, 2003; Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999).

The ANS is related to the two systems of “number sense” or non-symbolic number representation (Dehaene, 1997; 2001; 2009). One system appears to facilitate individuation, or the encoding of an exact cognitive representation for each item in a set (Hyde, 2011). The other system appears to be housed in the ANS and facilitates approximation, or encoding of an inexact representation for a set of objects (Hyde, 2011). More precisely, approximation may involve aggregation, or the encoding of information shared across set numbers (Hyde, 2011). Evidence for the individuation and approximation of non-symbolic number comes from two well-replicated findings. One, people individuate items for sets smaller than four (Feigenson et al., 2004; Scholl, 2001). For example, when shown a set of four dots, participants are more likely to remember
individual dots than the properties shared by the dots (Airely, 2001). Two, when encountering sets larger than four, people approximate items (Feigenson et al., 2004; Scholl, 2001). For example, when shown a set of eight dots, participants are more likely to erroneously recall a dot that represents the average dot size than an actual dot from the set (Airely, 2001). Similar evidence for the individuation and approximation of symbolic number comes from two recent studies. For sets of four numbers, adults were found to be more likely to individuate set numbers than aggregate across them (Cravalho et al., 2013; 2014). However, for number sets larger than four, adults were more likely to aggregate across set numbers (Cravalho et al., 2013; 2014).

The representational differences between the individuation and approximation systems have been proposed to be functional in nature (i.e., based on difference in strategy; Hyde, 2011). In terms of encoding strategies, the number of items in the set may hinder individuation, whereas approximation has no upper bound (Hyde, 2011). Hyde (2011) detailed evidence for these ideas regarding non-symbolic numerical processing, but the study by Brezis et al. (2015) provides evidence that they also apply to symbolic numerical processing. Other recent studies, discussed in detail below, provide additional evidence for the role of encoding strategies, working memory, and attention, in symbolic numerical processing. Attention is the selection of information for further processing as well as the blocking of other information from further processing (Anderson, 2004; Kane, Conway, Hambrick, & Engle, 2007). In the context of the current studies, more emphasis is put on the investigation of encoding and working memory capacity than on attention.
Strategies for Individual Number Set Processing

Multiple studies have used a task asking adults to first study a set of simultaneously presented 3-digit numbers and then immediately identify a member of the set (e.g., Cravalho et al., 2013; 2014). These studies were modeled on Airely’s (2001) non-symbolic individuation study that used dots instead of numbers. These number set studies found that, under differing task conditions, adults were able to successfully identify a set member at above chance levels from sets as large as eight numbers (Cravalho et al., 2013; 2014). However, as set size increased from four to six to eight numbers, participants more frequently misidentified the set mean as a member of the set (Cravalho et al., 2013; 2014). In another version of the task, modeled after Airely’s (2001) non-symbolic approximation study, participants were given limited time (less than .75 s) to study a set of numbers and then asked to identify the set mean (Cravalho et al., 2013). This study found that adults could successfully identify the set mean at above chance levels from sets as large as eight numbers (Cravalho et al., 2013). Self-report results suggested that representations and strategies for identifying a set member and the set mean differed.

When identifying the set member, self-reports indicated that adults scanned whole numbers, focused on the ones column, or tried to memorize the numbers (Cravalho et al., 2013; 2014). The results from experiments providing longer study and response times suggest that participants tried to scan and memorize individual numbers from sets of four and six numbers (Cravalho et al., 2013; 2014). Eye fixation data added detail to the self-reports. Specifically, adults looked most at the tens column regardless of set size, but as
set size increased they looked less at the tens column and shifted more attention to the ones column, which was the most diagnostic information for individuating these specific stimuli (Cravalho et al., 2014). In addition, study and response times were shorter for sets of eight, implying less memorization and more scanning of the ones column (Cravalho et al., 2013; 2014). Finally, in correspondence to the set size driven changes in eye fixations, study times, and response times, there was a positive relation between larger working memory capacity and more accurate identification of set members for sets of six numbers (Cravalho et al., 2013). This is consistent with the use of an individuation strategy in which maintaining individual numbers requires more working memory capacity for sets of six numbers than for sets of four numbers, as evidenced by a decline in accuracy as set size rose from four to six.

When identifying the set mean, self-reports indicated that adults also scanned whole numbers, but focused on the hundreds column instead of the ones column (Cravalho et al., 2013). Fixating on the hundreds column is a strategy consistent with adult number set comparisons (Morris & Masnick, 2015), for which existing eye tracking data shows that adults focus most on the hundreds column when comparing two sets of numbers. This is not surprising being that the task goal for identifying the set mean (i.e., to average the set numbers) would be similar to comparing sets to identify which set had the higher mean.

Overall, these studies of 3-digit number set processing provide evidence for successful encoding strategies for recognizing set members or set means. For number sets smaller than four, one is likely to individuate numbers via memorization and
focusing on the ones column to glean more diagnostic information about the set members than can be attained from the tens or hundreds columns. However, such individuation strategies might be hindered by our working memory capacity. For number sets larger than four, adults are likely to be successful by aggregating across set members, which reduces working memory processing. In summary, multiple recent studies support Hyde’s (2011) suggestion that different strategies are associated with different encoding goals and that strategy use is dependent on one’s working memory capacity (Brezis et al., 2015; Cravalho et al., 2013; 2014).

Strategies for Number Set Comparisons

When asking children to compare two sets of 3-digit numbers and infer which one has a higher mean, multiple studies have shown that children attend to between set mean ratio, as well as between and within set variability information when drawing their conclusions (Masnick & Morris, 2008; Morris et al., 2014). The results of these studies suggest that children adjust their processing strategy based on the mean and variance information that defined the presented sets (Masnick & Morris, 2008; Morris et al., 2014). Adults also used set mean and variance information in their 3-digit number set comparisons (Masnick & Morris, 2008; Morris & Masnick, 2015). Eye fixation data provided evidence that both children and adults focus more on the hundreds column of the numbers when the mean difference between sets is closer and when variance difference between sets is greater (Morris et al., 2014; Morris & Masnick, 2015). This is because when the goal is comparing means the hundreds column provides the most diagnostic information. These data suggest that children and adults process set
comparisons in a similar manner and both are quite accurate in their comparisons (Morris et al., 2014; Morris & Masnick, 2015). However, most likely due to greater knowledge of and more experience with number properties and comparisons (Morris et al., 2014), the efficiency of adult and child set comparison strategies differs considerably. Children often use less effective strategies, those of which are not employed by adults, like a subset comparison strategy in which they compare only the first two values of a set (Morris et al., 2014; Morris & Masnick, 2015). The use of this strategy may be related to why children are less likely to notice and use mean and variance information during set comparisons (Masnick & Morris, 2008).

Overall, children and adults who make accurate number set comparisons aggregated across the number set members and then compared the estimated average mean and variance values, producing accurate judgments of which sets had the higher means (Masnick & Morris, 2008; Morris et al., 2014; Morris & Masnick, 2015). However, children are less accurate and use more strategies than adults, some of which are less efficient (Masnick & Morris, 2008; Morris et al., 2014). This is most likely due to children having less knowledge and experience with number properties and comparisons than adults (Morris et al., 2014). To summarize, children and adults use similar strategies to compare 3-digit number sets, but children’s processing is less accurate and efficient.

**Literature Summary, Research Goals, and Hypotheses**

The current set of experiments was guided by the following research question, “Do adults and children infer the means and variances of number sets when making set
exemplar predictions?" The current experiments were guided by four goals: (1) to add to the literature of adult numerical predictions, (2) to extend the study of numerical predictions to children, (3) to investigate how working memory capacity influences the numerical predictions of adults and children, and (4) to examine the estimation strategies used by adults and children to generate numerical predictions. Strategies were measured via response times, self-reports, and eye fixation patterns.

For sets of simultaneously and sequentially presented numbers, multiples studies have shown that numerical mean judgments are accurate, but that judgment error increases with set variability (Anderson, 1964; 1968; Beach & Swenson, 1966; Lovie & Lovie, 1976; Spencer 1961; 1963). One previous number set prediction study that solicited numerical predictions found that participant’s predictions were, like number set mean judgments, close to the set means for sets of simultaneously presented numbers (Fowler, 1975). Other previous number set prediction studies have found that “deviant” set members, which create higher within-set variability, cause predictions to be further from the set means (Levin, 1976; Levin et al., 1977). In addition, there is evidence that both mean judgments (Malmi & Samson, 1983) and predictions (Pitz et al., 1976) are more indicative of set means than any other measure of central tendency (i.e., median, mode, midrange).

To add to the study of numerical predictions with adults, the first experiment of the current study followed Fowler (1975) and used sets of simultaneously presented numbers. Various hypotheses followed from the previous mean judgment and nominal prediction studies. First, it was hypothesized that the measure of central tendency that
adult and child number set predictions would reflect would be the set mean. This hypothesis was based on previous research and the idea that the arithmetic mean is the anticipated result of a continuous variable and therefore most suitable prediction for a future exemplar of that variable (Malmi & Samson, 1983; Nisbett & Ross, 1980).

Second, being that sets with high and low within-set variability were also used for the current prediction studies, it was hypothesized that predictions from sets with high set variability would be further from the set means than those from sets with low set variability. Multiple previous studies also supported that the framing of the mean judgment may prompt social bias that can result in the over- or underestimation of set means for both sets of simultaneously and sequentially presented numbers (Hendrick & Costantini, 1970; Levin 1974; 1975). Therefore, the prediction paradigm used for the current studies was designed to minimize such bias. One such precaution was to present the participants with a scenario in which the predicted outcome is trivial rather than socially relevant (e.g., IQ scores or store item prices; Levin 1974; 1975).

There are two major shortcomings of the prior number set judgment and prediction literature. One is that there are no studies that I could identify investigating children’s number set predictions. The third experiment of the current study extended the study of numerical predictions to 10-year-old children (4th grade students) who were subject to the American mathematics education standards that emphasize the representation and interpretation of number sets (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). It was hypothesized that they would infer set means and variances, just as adults did when making mean
judgments (Pollard, 1984) and predictions (Pitz et al., 1976) from number sets. However, it was also hypothesized that their predictions would produce values that were further from the set means than adult predictions because they should have less experience with numerical processing than adults (Masnick & Morris, 2008; Morris et al., 2014; Morris & Masnick, 2015).

The other major shortcoming of the existing literature is that, despite multiple studies discussing the role of a working memory dependent sampling strategy in number set processing (Brezis et al., 2015; Malmi & Samson, 1983; Lindskog et al., 2013a; 2013b; Lindskog & Winman, 2014; Spencer, 1963), none of these studies actually measured working memory capacity to investigate how it influenced the numerical predictions of adults and children. A related point is that of the previous studies only two involved participant processing of and numerical discernments from sets smaller than nine numbers (i.e., with each set number containing the same amount of digits). This is relevant because adult working memory processing is constrained to 4 ± 2 observations (Cowan, 2001) and children have even smaller working memory capacities (Davidson, Amso, Anderson, & Diamond, 2006; Gathercole, Pickering, Ambridge, & Wearing, 2004; Kail, 2007). In order to investigate how working memory capacity influences the numerical predictions of adults and children the current set of studies used smaller sets (e.g., sets of four) of only 3-digit numbers, namely sets of four, six, and eight numbers. Hence, set sizes within typical adult and child working memory capacities. Two prior studies of mean judgments have used set sizes within this range and illustrated how working memory capacity relates to strategy use. Approximate aggregation across
number set members was found to increase as set size increased from four to eight for both simultaneously (Cravalho et al., 2013) and sequentially (Brezis et al., 2015) presented numbers. Along with this increase, individuation of set members was found to decrease (Brezis et al., 2015; Cravalho et al., 2013). These findings supported Hyde’s (2011) assertions that approximation and individuation are strategic forms of numerical processing dependent on set size, and hence dependent on working memory. It was predicted that participants with larger working memory capacity should be able to maintain a greater number of items in working memory and more precisely estimate the means of the sets beyond typical working memory capacity (i.e., eight numbers).

Finally, the early study of numerical set processing practically ignored participant strategy, as Spencer’s (1963) study of numerical mean judgments was the only one to collect adult strategy self-reports. More recent studies have collected adult and child strategy self-reports for nominal number set mean judgments (Cravalho et al., 2013) and numerical number set mean comparisons (Masnick & Morris, 2008; Morris et al., 2014; Morris & Masnick, 2015), but no prior study had investigated strategies for number set predictions. Both children and adults aggregated across set members to compare number sets, but children used less efficient and effective encoding strategies, most likely due to less knowledge and experience with numerical processing (Masnick & Morris, 2008; Morris et al., 2014; Morris & Masnick, 2015). Being that mean judgments, comparing set means, and predictions share the same task goal of averaging set members (Levin, 1976; Levin et al., 1977; Peterson & Beach, 1967; Pitz et al., 1976; Pollard, 1984), it was hypothesized that strategies for making number set predictions would be similar to those
for mean judgments and set comparisons. For sets of simultaneously presented numbers, Spencer (1963) found that adults averaged samples, rather than try to process the entire set, to make their mean judgments. Also for sets of simultaneously presented numbers, multiple studies of number set comparisons found that children and adults focused on the 100s columns to average the 3-digit numbers because it is the most diagnostic information from which to make these comparisons (Morris et al., 2014; Morris & Masnick, 2015). Therefore, both adults and children were expected to average the sets of simultaneously presented numbers to make their predictions, focusing on the 100s column of the numbers or perhaps employing a sampling strategy to average a particular subset of set numbers, such as the minimum and maximum values (Spencer, 1963). It also followed that children’s prediction strategies would be less efficient than adult’s, similar to the pattern seen with number set comparisons (Masnick & Morris, 2008; Morris et al., 2014; Morris & Masnick, 2015). For instance, children often focused on non-strategic subsets during number set comparisons, regardless of whether those subsets contained useful information such as the minimum and maximum set values (Morris et al., 2014), so it was thought that a similar subset strategy would surface for children’s predictions, such as focusing on only the first two number in the set regardless of the set range.
CHAPTER III

METHODOLOGY

Overview of Experiments

Experiments 1 and 2 addressed three of the four goals outlined above in Chapter 2, specifically: (1) to add to the literature of adult numerical predictions, (2) to investigate how working memory capacity influences the numerical predictions of adults, and (3) to examine the estimation strategies used by adults to generate numerical predictions. The Experiment 1 procedure required adults to make their predictions from sets of simultaneously presented numbers. Experiment 2 replicated Experiment 1, as well as introduced a version of the task in which participants made their predictions from sets of sequentially presented numbers.

Experiment 3 addressed all four of the goals outlined above: (1) to extent the study of numerical predictions to children, (2) to add to the literature of adult numerical predictions, (3) to investigate how working memory capacity influences the numerical predictions of children, and (4) to examine the estimation strategies used by children to generate numerical predictions. Experiment 3 replicated Experiment 2 with adults and extended the paradigm to children. Participant strategy for processing the number sets was assessed via self-reports (all experiments), response times (Experiments 1 and 2; adults only), and eye fixation patterns (Experiment 3).

Experiment 1

The research question guiding Experiment 1 was, “Do adults infer the means and
variances of number sets when making set exemplar predictions from sets of simultaneously presented numbers?” A second research question specific to Experiment 1 was, “How does working memory capacity influence adult’s set exemplar predictions from sets of simultaneously presented numbers?” A final research question for Experiment 1 was, “How does the simultaneous presentation of a number set influence adult estimation strategies?”

Participants

Participants were undergraduate students recruited from the Kent State University Educational Psychology subject pool of participants. When looking for a medium effect with an estimated power of .75 at an alpha value of .05, the recommended number of participants for Experiment 1 was 35. This number was exceeded by a considerable amount, even though data from eight participants were identified as outliers (prediction averages were two standard deviations above the sample means) and not included in the analysis. The average age of the remaining participants included in the main analysis ($N = 81$) was $20.67$ ($SD = 3.98$), 74% were female, and most were of Caucasian descent. Two participants did not complete the ABCD grammatical reasoning task (cf. Was & Woltz, 2007), so they were excluded from the working memory analysis. For this reason, the average age of the participants included in the working memory analysis ($N = 79$) was $20.67$ ($SD = 4.03$), 73% were female, and most were of Caucasian descent. All participants were naïve to the purpose of the experiment and received course credit for their participation.
**Design**

Experiment 1 utilized a 3 (number set size) x 2 (number set variability) within-subjects design.

**Setting and Apparatus**

The study was conducted in the Educational Psychology Laboratory on the campus of Kent State University. This laboratory contained multiple computer stations. All experimental tasks were programmed in E-prime® and loaded onto the stations for individual participation.

**Materials**

The experimenter created a number set exemplar prediction task, which was used to measure participant exemplar predictions, response latency, and strategy self-reports (frequency rating and open-ended). The ABCD grammatical reasoning task (cf. Was & Woltz, 2007) was used as a measure of verbal working memory capacity.

**Number set exemplar prediction task.** The number set exemplar prediction task required participants to study a set of numbers, each one representing how far a batter hit a baseball, and then predict how far the batter would hit the next baseball. This task consisted of 96 trials. Each trial contained three parts, described here in order of presentation (see Figure 1, Slides 1-3). First, a fixation cross was presented for 500ms. Then, a set of 3-digit numbers was shown for one second per set number (details of the set sizes and presentation durations below). Showing each set number for one second is a relatively long amount of time compared to previous number set research that utilized presentation times shorter than a second (Anderson, 1968; Beach & Swenson, 1966;
Brezis et al., 2015; Cravalho et al., 2013; 2014; Levin, 1974; 1975; Malmi & Samson, 1983; Spencer, 1961; 1963). However, this longer presentation time was chosen to remain consistent throughout the experiment, unlike the previous experiments, because shorter presentation rates make it difficult to successfully use mathematical processing strategies, and are associated with a systematic underestimation of the mean (Malmi & Samson, 1983). Therefore, using a shorter presentation time would have defeated the main purpose of the study, which was to investigate participant strategies. Finally, the participant was prompted to type in how far they thought the batter would hit the next baseball, with a blinking cursor indicating where they would see their typed response. Responses were limited to four digits. After typing in their response, they were prompted to press the spacebar to continue to the next trial. All slides for the exemplar prediction task utilized black symbols against a white background.
Instructions preceded each of four blocks of number sets. First participants made predictions for a block of six practice trials that were not analyzed. Experimental trials included the three following blocks: (a) 30 sets of four numbers (each set presented for 4000ms), (b) 30 sets of six numbers (each set presented for 6000ms), and (c) 30 sets of eight numbers (each set presented for 8000ms). The blocks were presented in the same order as listed. Within each experimental trial block, half of the sets were drawn from one of two variance types, either 10% (low variance; e.g., 345, 404, 367, and 308) or 20% (high variance; e.g., 484, 608, 683, and 409) of the set mean. All participants saw the same practice trials presented in the same order. However, for the experimental trials, the presentation order was randomized within each block.
After completing each block of experimental trials, participants were asked to report on their strategy use. First, participants were presented with the following strategy descriptions (cf. Masnick & Morris, 2008): (a) “Look at the first digit (e.g., the 1 in the 125).”, (b) “Look at the second digit (e.g., the 2 in 125).”, (c) “Look at the third digit (e.g., the 5 in 125).”, (d) “Try to figure out the average.”, (e) “Find the biggest number.”, (f) “Find the smallest number.”, (g) “Just get a sense of the numbers.”, (h) “Look for a number that is not like other numbers.”, and (i) “Try to memorize specific numbers.”. Participants rated how often they used each strategy by pressing a number key corresponding to the following scale: 1) never, 2) some trials, 3) most trials, or 4) always. Each strategy description and the response options remained on the screen until the participant chose an answer. After completing the frequency-rating portion of the strategy report, participants were asked to type a brief description of the strategy they used for predicting how far the batter would hit the next baseball, with a blinking cursor indicating where they would see their typed response (see Figure 1, Slide 4). After typing in their response, they were prompted to press the nine key to continue on to the next experimental trial block.

**ABCD working memory task.** The verbal working memory task required participants to identify the order of letter pairs from statements they were previously shown. This particular task involves remembering simple and challenging statements regarding the order of letters (the ABCD grammatical reasoning task: cf. Was & Woltz, 2007). The simple statements include two letters (e.g., “D comes after C”) and the challenging statements include four letters and a set designation (e.g., “A comes after B,
C comes before D, Set 1 comes after Set 2”). The set designation is that Set 1 always refers to the A and B order and that Set 2 always refers to the C and D order.

First participants completed four practice simple statement trials, then 16 experimental simple statement trials. The practice trials were prefaced with detailed instructions and an example of what the response screen would look like. Each simple statement trial adhered to the following order (see Figure 2). The words “Get Ready” were shown in the middle of the screen for 1000ms, followed by an asterisk (250ms). Then a statement was shown in the middle of the screen along with response options (e.g., “CD”, “DC”) in the lower left and lower right part of the screen. Participants were prompted to press the one key to choose the response on the lower left and the two key to choose the response on the lower right. Participants received feedback of “Correct!” or “Incorrect” for each response. On the experimental trials, response latency was also displayed for correct responses. More detailed feedback was given during the practice trials (e.g., “D comes after C = DC”) regardless of whether the response was correct or incorrect.
Next participants completed two practice challenging statement trials, then 24 experimental challenging statement trials. The practice trials were prefaced with detailed instructions and two examples. Each challenging statement trial adhered to the following order (see Figure 3). The words “Get Ready” were shown in the middle of the screen for 1000ms, followed by an asterisk (500ms). Then, three sequences were displayed in random order. One sequence begins with a letter displayed for 1500ms, then the phrase “comes after” displayed for 1500ms, followed by another letter displayed for 1500ms. This sequence was used twice for each trial, one time with the letters A and B and one time with the letters C and D. The other sequence begins with a set designation displayed for 1500ms, then the phrase “comes after” displayed for 1500ms, followed by the other set designation displayed for 1500ms. Each sequence ends with the phrase “Press the spacebar when you are ready to continue.” being displayed for 30000ms. Then
participants were prompted to select one of eight answer options (e.g., “ABCD”) using the one through eight keys. Participants received feedback of “Correct!” or “Incorrect” for each response. After finishing the experimental challenging statement trials, participants completed 16 more experimental simple statement trials. All slides for the verbal working memory task utilized black symbols against a white background.
Figure 3. Example of an ABCD grammatical reasoning task challenging statement trial.

Procedure

Each participant was seated at one of four computer stations and informed consent was obtained. Then the participant completed the number set exemplar prediction task
and the ABCD working memory task using the keyboard at the computer station. The order of the two tasks was counterbalanced across participants. After completing both tasks, the participant was free to leave the lab.

**Hypotheses**

When asked to predict future set exemplars, it was expected that participants would generate values close to the set means. This was expected because prior research found that adult’s numerical predictions aligned closely with the set means (Fowler, 1975; Pitz et al., 1976). However, it was expected that predictions would be closer to the set means for the larger sets (i.e. six & eight numbers) than for sets of four numbers. This was expected because given sets larger than four, people approximate items, (i.e., average over set members retaining information about set features; Scholl, 2001), whereas given sets smaller than four, they individuate items (i.e., represent individual set members; Scholl & Leslie, 1999). Brezis et al. (2015) provided evidence that number set processing is dependent on these same patterns, with sets of four numbers being encoded via individual number representations and sets of eight numbers being encoded via set aggregates. Therefore, participants were expected to individually represent numbers from sets of four, which would interfere with their approximations, but such interference was not expected when participants represented sets of six or eight numbers. Further, it was hypothesized that exemplar predictions for number sets with lower set variability would be closer to the set means than predictions for number sets with higher set variability. Based on prior number set mean estimation studies, estimations error increases with variance (Beach & Swenson, 1966; Levin, 1975; Spencer, 1961; 1963), so
it was expected that a similar effect would be present with numbers set predictions. In addition, predictions for sets with lower within-group variability were expected to be given faster because they would seem to enable more efficient encoding. For example, the following low variability set, 370, 391, 340, and 319, may only require one scan of the hundreds column and the participant to remember that all the numbers begin with 300, whereas the following high variability set, 721, 645, 475, and 551, may require multiple scans of the hundreds column and the participant to remember that the four numbers all begin with different hundreds values.

Although no previous studies have investigated the role of working memory in numerical set predictions, it was thought that higher verbal working memory capacity would be related to exemplar predictions, but only for sets of eight numbers. This was expected because normal adult WORKING MEMORY capacity (4 ± 2 observations; Cowan, 2001) provides enough capacity to process sets of four and six numbers (Brezis et al., 2015), but not enough for set of eight numbers.

Finally, it was expected that participants would report approximation strategies similar to those from previous experiments including approximation tasks (Cravalho et al., 2013; Morris et al., 2014). For example, looking at the 100s column of the number sets implied that one made use of the most diagnostic information for approximating overall set characteristics (Cravalho et al., 2013; 2014; Morris et al., 2014). Therefore, looking at the 100s column of the number sets presented in Experiment 1 was an expected strategy.
Experiment 2

Experiment 1 provided evidence that adults infer the means and variances of numbers sets when estimating predictions from those sets and that those inferences influence their estimation strategies. However, the design of the Experiment 1 exemplar prediction task may have limited the processing strategies used by the participants, as they were only allowed to view sets of simultaneously presented numbers. Experiment 2 provided a comparison between predictions from sets of simultaneously presented numbers (i.e., all the set numbers presented at once) and sets of sequentially presented numbers (i.e., one set number presented at a time). Therefore, the first research question for Experiment 2 was, “Do adults infer the means and variances of number sets when making set exemplar predictions from sets of sequentially presented numbers?”

The additional set presentation type also changed the task demands. For instance, simultaneous presentation prompts one to shift their attention between numbers, though one does not need to hold any numbers in working memory. However, during sequential presentation participants will need to hold set numbers that were previously displayed in working memory and update their prediction based on how those numbers relate to the current number they are encoding. Therefore, the sets of sequentially presented numbers should provoke different processing strategies than those for the sets of simultaneously presented numbers. Hence, two additional research questions were addressed. The second research question was, “How does working memory capacity influence adult’s set exemplar predictions from sets of sequentially presented numbers?” The final research
question was, “How does the sequential presentation of number sets influence adult estimation strategies?”

In addition, the inclusions of the description “try to figure out the average” in the Experiment 1 strategy frequency ratings may have had an unintended influence on the open-ended strategy reports and the strategies used. Specifically, this description may have alerted the participants to the strategy of trying to figure out the average for each set, causing more participants to report this strategy in their open-ended strategy descriptions than may have otherwise. This being a potential confound, the frequency-rating portion of the strategy self-report was removed from Experiment 2.

Participants

Participants were undergraduate students recruited from the Kent State University Educational Psychology course pool of participants. For the version of the prediction task displaying set numbers simultaneously, data from two participants were identified as outliers (prediction averages were two standard deviations above the sample means) and were excluded from the analysis. The average age of the participants (N = 35) included in the main analysis and working memory analysis for this version of the task was 19.77 (SD = 1.06), 80% were female, and most were of Caucasian descent. For the version of the prediction task displaying set numbers sequentially, data from six participants were identified as outliers (prediction averages were two standard deviations above the sample means) and were excluded from the analysis. Based on the previously discussed power analysis, the elimination of these outliers left the main and working memory analyses for this version of the task underpowered. The average age of the remaining participants (N
was 19.74 (SD = .93), 77% were female, and most were of Caucasian descent. All participants were naïve to the purpose of the experiment and received course credit for their participation.

**Design**

Experiment 2 utilized a 3 (number set size) x 2 (set variability) x 2 (set presentation) within-subjects design.

**Setting and Apparatus**

The study took place in the same location as Experiment 1. All experimental tasks were programmed in E-prime®.

**Materials**

In addition to a modified version of the simultaneous presentation version of the number set exemplar prediction task used in Experiment 1, a sequential presentation version of the exemplar prediction task was created. Each task measured exemplar predictions, response latency, and open-ended strategy self-reports. The ABCD grammatical reasoning task (cf. Was & Woltz, 2007) was again used to measure participant verbal working memory capacity.

**Number set exemplar prediction task.** Each version of the prediction task followed the same procedure as the Experiment 1 prediction task except for the following. In one version, the numbers in each set were presented simultaneously just as in Experiment 1. However, there were only three practice trials and 20 experimental trials per block and participants were only asked to type a brief description of their strategy for each block. For the other version, the numbers in each set were presented
sequentially. Each sequential trial adhered to the following pattern (see Figure 4). A fixation cross was presented for 500ms before each number in the set was shown. No matter the size of the set, each number in the set was shown for 1000ms. As in the simultaneous presentation version, there were only three practice trials and 20 experimental trials per block and participants were only asked to type a brief description of their strategy for each block. Fewer trials were used per block in order to accommodate the within-subjects design of the task, allowing each participant to complete both versions of the prediction task in one session. In addition, each participant completed the same verbal working memory task used in Experiment 1.
Procedure

Each participant was seated at one of four computer stations and informed consent was obtained. Then the participant completed both versions of the number set exemplar prediction task and the ABCD working memory task using the keyboard at the computer station. The order of the exemplar prediction tasks was counterbalanced across participants, with the ABCD working memory task always being completed between the
prediction tasks. After completing all three tasks, the participant was free to leave the lab.

**Hypotheses**

Following the results of Experiment 1, it was predicted that, for sets of simultaneously presented numbers, exemplar predictions would be closest to the mean for sets of eight numbers, which would statistically differ from predictions for sets of four, but not from sets of six. Also for sets of simultaneously presented numbers, it was expected that predictions for sets of four would be underestimated, whereas predictions for sets of six and eight would be overestimated. In addition, no differences in predictions for high variability and low variability sets were expected. Finally, response times were expected to decrease as set size increases, with low variability responses times being faster than high variability response times only for sets of eight numbers.

For sets of sequentially presented numbers, it was predicted that exemplar predictions would also be closest to the mean for sets of eight numbers. Being that presentation time per number and response format are uniform across presentation types, approximation strategy use for sequentially presented numbers should be most prevalent for sets of eight as it was for simultaneously presented numbers. There was also no difference expected between predictions for high variability and low variability sets, again because the timings are equal across presentation type.

Although no previous studies conducted within-subject comparisons of numerical set predictions from sets of simultaneously and sequentially presented numbers, it was predicted that participants set exemplar predictions would be closer to the set means
when the numbers were presented simultaneously than when presented sequentially. This was expected because, unlike sequential presentation, simultaneous presentation would not require participants to hold any set numbers in working memory or to update their prediction based on how the previously encoded numbers relate to a newly encoded number. Therefore, approximation processing should be more precise (i.e., closer to the mean) when numbers are presented all at once. Also due to the additional working memory and encoding constrains of processing one number at a time, response times were expected to be slower for predictions from sequentially presented numbers than for simultaneously presented numbers.

Following the results of Experiment 1, higher verbal working memory capacity was not expected to be related to predictions from sets of simultaneously presented numbers. However, higher verbal working memory capacity was expected to be related to exemplar predictions from sets of sequentially presented numbers. Specifically, being forced to use more complicated strategies, due to the additional attention and encoding constrains of processing one number at a time, was thought to require more capacity for a precise estimation. Due to the increase in processing constrains discussed above, and the limits of normal adult working memory capacity (4 ± 2 observations; Cowan, 2001), it would seem participants only have enough capacity to process sets of four numbers, but may have trouble with six numbers, and most likely not enough capacity for sets of eight numbers (Brezis et al., 2015).

Finally, although no previous studies have compared strategies for processing sets of simultaneously and sequentially presented numbers, it was expected that the
presentation types would provoke different strategies. For instance, participants from Experiment 1 (simultaneous presentation) reported looking at the 100s column for all set sizes, but more so for sets of six and eight numbers. That strategy seemed to have helped them to approximately average the set numbers, which they also reported doing, a notion that is supported by their predictions being close to the set means. However, when numbers are presented sequentially, participants may also try to round the numbers (Brezis et al., 2015). For example, rounding to the 100s column may make it easier to hold previously seen set numbers in working memory as one updates their prediction based on the number they are currently encoding. This may cause them to report paying more attention to the 10s column (i.e., in order to round to the 100s column) when the numbers are presented sequentially than when they are presented simultaneously.

**Experiment 3**

Experiment 2 provided more evidence that adults infer the means and variances of numbers sets when estimating predictions from those sets, showing that they do so whether the set numbers are presented simultaneously or sequentially. Although there were not many differences in the strategies used, Experiment 2 also showed how inferences of set means and variances influenced adult estimation strategies when the numbers are presented in different ways. However, neither Experiment 1, nor Experiment 2 provided objective evidence for participant strategies, as strategies were only inferred from participant self-reports. Experiment 3 provided similar experimental conditions as the previous experiments, but also included the collection of eye tracking data as a dependent measure. The addition of these data was thought to allow more direct
insight into participant strategy use by providing direct evidence for their attention to specific number set information during encoding. Adult strategy self-reports were not collected for Experiment 3 as it was intended to triangulate the Experiment 3 adult eye fixation data with the adult self-reported strategies from Experiments 1 and 2. Another change from Experiments 1 and 2 was that the adult participants did not complete the ABCD working memory task (cf. Was & Woltz, 2007). This was because the focus of Experiment 3 was to investigate attention, rather than working memory, and number set predictions.

Experiment 3 also served as a replication of the first two experiments. Being that the prediction results from sets of simultaneously presented numbers from Experiment 1 were only partially replicated by Experiment 2, the third experiment was meant to clarify the discrepancy in findings. It was thought that the Experiment 1 results may be skewed due to the larger sample of participants and that the Experiment 2 results, which were based on the exact recommended amount of participants from a power analysis, were actually more indicative of predictions from sets of simultaneously presented numbers. In addition, Experiment 3 served as a replication of the adult prediction results from sets of sequentially presented numbers generated from Experiment 2. Finally, Experiment 3 extended the exemplar prediction paradigm to children in order to compare their performance and strategies to that of adults.

The general methodology of the exemplar prediction tasks from Experiment 2 were left intact for Experiment 3 since it was not known how the differences in number set presentation would effect children’s predictions. Initially, as with the adults, strategy
self reports were not going to be collected from the child participants, but once it was apparent that children may report using different strategies than the adults, open-ended strategy self-reports began being collected after a child had completed each version of the exemplar prediction task. Also unlike the adult participants, the child participants did complete a working memory task, the automated working memory assessment (AWMA; c.f. Alloway, Gathercole, Kirkwood, & Elliott, 2008). A control task was also added to ensure that the children exhibited mature numerical representations, which if lacking can cause a lag in the development of other math skills (Siegler, Thompson, & Opfer, 2009). This task was magnitude estimation via 0-1,000 number lines, which is the range of the number set stimuli that was used for Experiment 3.

The first research question for Experiment 3 was, “Do children infer the means and variances of number sets when making set exemplar predictions from sets of simultaneously and sequentially presented numbers?” The second research question was, “How does working memory capacity influence children’s set exemplar predictions from sets of simultaneously and sequentially presented numbers?” The third research question was, “How does the simultaneous and sequential presentation of number sets influence child estimation strategies?” The final research question was, “What set aspects do adults and children pay attention to when estimating predictions from differently presented number sets?”

**Adult Participants**

The adult participants were undergraduate students recruited from the Kent State University Educational Psychology course pool of participants. For the version of the
prediction task displaying set numbers simultaneously, data from four participants were identified as outliers (prediction averages were two standard deviations above the sample means) and were excluded from the analysis. The average age of the participants \((N = 36)\) included in the main analysis and working memory analysis for this version of the task was 19.67 \((SD = .96)\), 81% were female, and most were of Caucasian descent. For the version of the prediction task displaying set numbers sequentially, data from a differing combination of four participants were identified as outliers (prediction averages were two standard deviations above the sample means) and were excluded from the main and working memory analyses. The average age of the remaining participants \((N = 36)\) was 19.75 \((SD = 1.03)\), 81% were female, and most were of Caucasian descent. All participants were naïve to the purpose of the experiment and received course credit for their participation.

**Child Participants**

The child participants were 4th grade students recruited from local schools with assistance from the Research Center for Educational Technology’s AT&T Classroom (on the campus of Kent State University). In total, 43 children provided exemplar prediction data. However, due to a design confound with the sequential version of the exemplar prediction task (see limitations section of discussion for details), eight participants were found to be adding the set numbers presented one at a time. Data from these eight participants were removed from both the simultaneous and sequential task analyses. The average age of the participants included in the main analyses \((N = 35)\) was 10.25 \((SD = .34)\), 49% were female, and most were of Caucasian descent.
Although the majority of the child participants completed the short version of the automated working memory assessment (AWMA; c.f. Alloway et al., 2008), the initial 10 children completed only the exemplar prediction tasks and the number line task before the experimenter gained access to the AWMA. For this reason, the average age of the participants included in the working memory analyses \((N = 28)\) was 10.25 \((SD = .36)\), 54% were female, and most were of Caucasian descent. Based on the previously discussed power analysis, this number of participants left the child working memory analyses underpowered. All child participants were naïve to the purpose of the experiment and received stickers or a pencil for their participation.

Forth graders were used because most children at this level of schooling have a more precise representation of numbers in the 0-1,000 range. Specifically, if asked to estimate where numbers fall on a 0-1,000 number line, they have a linear representation, evenly spacing the numbers across the number line, rather than a compressed logarithmic representation, where estimates for smaller numbers tend to be disproportionately closer to the “0” mark on the line then they should be (Siegler et al., 2009). The occurrence of a logarithmic representation would have been rare for the undergraduates and 4th graders tested given that some 2nd graders’ number line estimations reflect a linear representation of numbers within the 0-1,000 range (Siegler et al., 2009).

**Design**

Experiment 3 utilized a 3 (number set size) x 2 (set variability) x 2 (set presentation) within-subjects design.
**Adult Participant Setting and Apparatus**

The adult phase of the study took place in the Educational Psychology Laboratory on the campus of Kent State University. This laboratory contains a larger station partitioned off from the computer stations used to run Experiments 1 and 2. All experimental tasks were programmed using Tobii® software and projected through a Tobii® T-60XL eye tracker connected to a laptop, except for the magnitude estimation task which was completed via paper number lines and pencil markings (see details below). It is important to note that the Tobii® software and hardware used for the exemplar prediction tasks did not allow for the collection of response time data, so this was not a dependent variable as it was in Experiments 1 and 2. Participants completed all experimental tasks individually.

**Child Participant Setting and Apparatus**

The child phase of the study took place in multiple locations, either at the AT&T classroom (on the campus of Kent State University) or in the school from which the children were recruited. In the space provided, the Tobii® T-60XL eye tracker and laptop were set up for data collection using the same tasks completed by the adults. However, child participants also completed a verbal working memory task on a separate laptop. The child participants also completed magnitude estimations via the same paper and pencil number line task completed by the adult participants.

**Materials**

Modified versions of the simultaneous presentation and sequential presentation exemplar prediction tasks were created for Experiment 3. Each task measured exemplar
predictions and eye fixations. The child participants were also asked to verbally report on their number set processing strategies and to complete the short version of the AWMA, which measured their verbal working memory capacity. In addition, each adult and child completed a number-to-position magnitude estimation task measuring his or her mental representation of numbers in the 0-1,000 range.

**Number set exemplar prediction task.** The two number sets prediction tasks followed the same procedure as in Experiment 2 except for the following changes to both tasks. Experiment 3 included the collection of eye fixation data via a Tobii® T-60XL eye tracking monitor. Participants were seated approximately 70 cm from the monitor and a nine-point calibration was performed with each participant prior to beginning each exemplar prediction task. Rather than reading instructions on the monitor screen, the researcher verbally explained the task before the participant began the practice trials. Each set of numbers was presented, via a PowerPoint® slide, in 48-point Times New Roman font with two spaces between the hundreds and tens columns and two spaces between the tens and ones columns. There were only 14 experimental trials per block and participants verbally reported their predictions to the experimenter, who wrote them down on a sheet delineating each block and trial. Each participant saw the trials in the same order, as the Tobii® software available to the experimenter did not allow for randomization of the trials. Only child participants were asked to report on their strategy use via an open-ended prompt. The experimenter took notes on their verbalized strategy explanations.
**Number line task.** The number-to-position magnitude estimation task used number lines indicating (at the top of the page) a number whose position was to be estimated on a line (in the middle of the page) marked with “0” on the left and with “1,000” on the right (see Figure 5; Thompson & Opfer, 2008). Participants marked the location of 22 numerals (i.e., 2, 5, 18, 34, 56, 78, 100, 122, 147, 150, 163, 179, 246, 366, 486, 606, 722, 725, 738, 754, 818, 938) by drawing a hatch marks on separate pages with 20-cm number lines where they thought the numerals were located (cf. Thompson & Opfer, 2008). Using these numerals maximized the discriminability of logarithmic and linear functions and minimized the influence of orientation strategies (e.g., using the knowledge that 500 is halfway between 0 and 1,000; Thompson & Opfer, 2008).

*Figure 5.* Example of an unmarked 0-1,000 number line from the number-to-position magnitude estimation task.
Automated Working Memory Assessment. The child participants also completed the short version of the AWMA (c.f. Alloway et al., 2008). This is the computerized version of the working memory test battery for children (WMTB-C; Pickering & Gathercole, 2001), which assesses verbal working memory. From this assessment, children’s standardized digit recall scores were used for statistical analysis. The digit recall task consists of an automated narrator first saying strings of numbers and the child then repeating the numbers. There are three practice trials, first repeating one number, then two numbers in a row, and finally three numbers in a row. For the proper task, trials are broken into blocks based on how many numbers are to be repeated back, beginning with repeating one number and going up to as many as nine numbers. There are six possible trials per each block, but once a child completes four trials (correctly repeating back all of the numbers within each narrated string), the narrator would move on to the next block. If for any block a child were unable to complete four trials, the task would end.

Procedure

Each adult participant was seated at a larger computer station that was partitioned off from the computer stations used to run Experiments 1 and 2 and informed consent was obtained. Then the participant completed both versions of the number set exemplar prediction task and the number-to-position magnitude estimation task. When possible, the order of the exemplar prediction tasks was counterbalanced across participants, with the number-to-position magnitude estimation task always being completed between the
prediction tasks. After completing all three tasks, the participant was free to leave the lab.

Data collection with the children was divided into two sessions. At the beginning of each session, the participant was seated at a table and informed consent (assent) was obtained. Then the participant completed one of the versions of the number set exemplar prediction task and either the AWMA or the number-to-position magnitude estimation task. During the second session, the child completed the other version of the exemplar prediction task and whichever of the AWMA or number line task they had not already completed. When possible, the order of the exemplar prediction tasks, as well as the AWMA and number tasks, was counterbalanced across participants. After completing the two tasks for that session, the participant was free to return to class.

**Adult Participant Hypotheses**

Following the results of Experiments 1 and 2, it was predicted that, for sets of simultaneously presented numbers, exemplar predictions would be furthest from the mean for sets of six numbers, which would statistically differ from predictions for sets of four numbers, but not from sets of eight numbers. Also for sets of simultaneously presented numbers, it was expected that predictions for sets of four would be underestimated, whereas predictions for sets of six and eight would be overestimated. In addition, no differences between predictions for high variability and low variability sets were expected.

Following the results of Experiment 2, it was predicted that, for sets of sequentially presented numbers, exemplar predictions would not differ based on set size.
It was also expected that predictions for each set size would be underestimated. No difference was expected between predictions for high variability and low variability sets.

Again following the Experiment 2 results, it was predicted that participants set exemplar predictions would be closer to the set means when the number sets were presented sequentially, but only for sets of six numbers. In addition, it was expected that the eye tracking data would reflect the previously reported adult strategies of looking for the biggest and smallest numbers for simultaneously presented numbers and looking at the 100s column for both simultaneously and sequentially presented numbers.

**Child Participant Hypotheses**

No previous studies had investigated children’s number set predictions and processing strategies, so the following hypotheses were all conservative. Due to having less experience processing sets of numbers than adults, it was predicted that children’s predictions would be further from the set means for both simultaneously and sequentially presented numbers. It was expected that children’s performance and strategy use would differ because they have smaller attention spans and working memory capacities than adults (Astle & Scerif, 2009; Davidson et al., 2006; Gathercole et al., 2004; Kail, 2007; Ruff & Lawson, 1990). Even though working memory capacity was not found to play a role in adult number set predictions, it was thought that since children have less attentional and working memory capacity, and would most likely use less effective strategies, that working memory capacity would play a role in their predictions. It was predicted that children with higher capacity would make predictions closer to the set
means of both simultaneously and sequentially presented numbers than those with lower capacity.

Previous research had demonstrated that, like adults, children attend to set means and variances when making comparisons between number sets, although they employ processing strategies that are less efficient and effective than adults (Masnick & Morris, 2015; Morris & Masnick, 2008; Morris et al., 2014). Number set comparisons provide a similar task goal (i.e., to average the set numbers) as estimating a prediction from a number set, so it was expected that children would infer set means and variances when making number set exemplar predictions, but would use less efficient and effective strategies than those exhibited by the adult participants in Experiments 1 and 2. For example, adults pay most attention to the 100s column of the set numbers and pay little attention to the 1s column, but children may pay more equal attention to the 100s and 1 columns.
CHAPTER IV
ANALYSIS OF THE FINDINGS

Experiment 1

Experiment 1 added to the literature on adult numerical predictions from sets of simultaneously presented numbers, as well as to the literature on strategies used by adults to generate numerical predictions from sets of simultaneously presented numbers. Experiment 1 was also the first experiment to investigate how working memory capacity influences adult numerical predictions from sets of simultaneously presented numbers.

Exemplar Predictions

Participant predictions were standardized via conversion to z-scores before being analyzed. All figures display prediction z-scores in order to clearly illustrate how far predictions were from the set means.

Descriptive overview. Participant predictions were standardized via conversion to z-scores before being analyzed. All figures display prediction z-scores in order to clearly illustrate how far predictions were from the set means. Predictions for sets of four and sets of six were, on average, within 7% of the set means (see Figure 6), but predictions were within 1% of the set means for sets of eight numbers. Participant predictions underestimated the set means for set of four numbers, but overestimated the set means for sets of six and eight numbers.

When comparing predictions from sets with high and low set variability, there was no difference for sets of four numbers, but the latter were closer to the set means for
sets of six numbers and further from the set means for sets of eight numbers. Finally, the between set size changes in how close predictions were to the set means were also more drastic for the high variability sets than for the low variability sets.

**Figure 6.** Mean Experiment 1 prediction z-scores for sets of simultaneously presented numbers by set size and set variability. *Note.* Error bars represent SE of mean.

**Set size and set variability findings.** A 3 (set size) x 2 (set variability) repeated measures ANOVA was used to analyze the exemplar prediction data. Set size had a significant effect on how close predictions were to the set means, $F(2, 160) = 10.68, p < .001, \eta^2 = .12$. Within-subjects contrasts showed that predictions for sets of four and sets of six were significantly different, $F(1, 80) = 21.32, p < .001, \eta^2 = .21$, although this reflects the underestimation of sets of four and the overestimation of sets of six, rather than a difference in how close the predictions were from the set means. In fact, predictions for sets of four and sets of six were the exact same distance from the set means (i.e., 7%). Within-subject contrasts also showed no difference between
predictions for sets of six and sets of eight, $F(1, 80) = 3.71, p = .06, \eta^2 = .04$. A paired-samples t-test indicated a significant difference between predictions for sets of four ($M = -.07$) and sets of eight ($M = .01$), $t(80) = -2.67, p < .01$, 95% CI [-.13, -.02], $d = .30$.

These results only partially supported the hypothesis that predictions for larger sets (i.e., six and eight numbers) would be closer to the set means than predictions for sets of four numbers, as only predictions for sets of eight were significantly closer to the set means than predictions for sets of four.

There was not a significant effect of set variability on predictions, $F < 1$. This finding did not support the hypothesis that predictions for low variability sets would be closer to the means than those for high variability sets. There was not a significant interaction between set size and set variability, $F(2, 160) = 1.41, p = .25, \eta^2 = .01$.

**Working memory findings.** To test the hypothesis that participants with more working memory (WM) capacity would make predictions that are closer to the actual set means for sets of eight numbers, additional analyses were conducted. To this end, a quartile split was conducted to separate the participants into groups based on highest ($N = 19$), high ($N = 20$), low ($N = 20$), and lowest ($N = 20$) verbal WM capacity. Then a one-way ANOVA, with working memory capacity groups as a between subjects factor, was performed. WM group was not found to have an overall effect on participant predictions for sets of eight numbers, $F < 1$, and Bonferroni corrections stated that comparisons among the four WM groups yielded no significant differences in how close predictions were to the set means. These results did not support the researcher’s hypothesis.
Response Times

When the participant was prompted to type in how far they thought the batter would hit the next baseball, the time it took to provide a prediction was measured in milliseconds, thus all figures display participant response times in milliseconds.

Descriptive overview. The time it took to produce a prediction steadily decreased as set size increased (see Figure 7). Across set size, when comparing response times from sets with high and low set variability, the latter were faster than the former.

Figure 7. Mean Experiment 1 response times for sets of simultaneously presented numbers by set size and set variability. Note. Error bars represent SE of mean.

Set size and set variability findings. A 3 (set size) x 2 (set variability) repeated measures ANOVA was used to analyze the response time data. Set size did have a significant effect on response times, \( F(2, 160) = 12.05, p < .001, \eta^2 = .13 \). Within-subjects contrasts showed that there was a significant decrease in response times as set size increased from four to six, \( F(1, 80) = 7.45, p < .01, \eta^2 = .09 \), and as set size
increased from six to eight, $F(1, 80) = 7.50, p < .01, \eta^2 = .09$. A paired-samples t-test also indicated a significant difference between response times for sets of four ($M = 3003.82$) and sets of eight ($M = 2555.21$), $t(80) = 4.26, p < .001, 95\%$ CI $[238.80, 658.41], d = .45$.

Set variability also had a significant effect on response time, $F(1, 80) = 6.98, p < .05, \eta^2 = .08$. Paired samples t-tests, with effect size corrections for dependence between means (Morris & DeShon, 2002), were conducted to further analyze the effect of set variability on response times per each set size. For sets of four, response times for low variability sets ($M = 2959.76$) were not significantly faster than response times for high variability sets ($M = 3053.82$), $t(80) = 1.21, p = .23, 95\%$ CI $[-60.43, 248.54], d = .14$. For sets of six, response times for low variability sets ($M = 2710.33$) were not significantly faster than response times for high variability sets ($M = 2825.58$), $t(80) = 1.72, p = .09, 95\%$ CI $[-18.27, 248.76], d = .22$. However, for sets of eight, response times for low variability sets ($M = 2504.30$) were significantly faster than response times for high variability sets ($M = 2618.07$), $t(80) = 2.26, p < .05, 95\%$ CI $[13.37, 214.17], d = .27$. These findings partially supported the hypothesis that predictions for low variability sets would be given faster than predictions given for high variability sets. There was not a significant interaction between set size and set variability, $F < 1$.

**Strategy Self-reports**

Participants provided self-reports on their strategy use after each set size block of prediction task trials.
**Summary of frequency ratings.** A segment of the frequency ratings for the strategy descriptions are listed here in descending order of usage (see Table 1): (a) “just get a sense of the numbers”, (b) “find the biggest number”, (c) “find the smallest number”, (d) “try to figure out the average”, (e) “look for a number that is not like other numbers”, and (f) “try to memorize specific numbers”. Participants’ reported use of these strategies remained fairly stable as set size increased, with paired samples t-tests finding no significant differences between frequency ratings among the set sizes. Of more interest was the participants’ reported attention to the specific digit columns of the number sets (see Table 1). Even though participants paid the most attention to the 1st digit column (i.e., the 100s column of the number sets) across set size, they focused even more so on the 100s column as set size increased from four ($M = 3.12$) to six ($M = 3.49$), $t(88) = -3.56, p < .01, 95\%\ CI [-.58, -.16], d = .38$. The high rate of attention to the 100s column was sustained as set size increased from six to eight ($M = 3.49$), $t < 1$. Regardless of set size, the same amount of attention was paid to the 2nd digit column (i.e., the 10s column of the number sets), with paired samples t-tests finding no significant differences between frequency ratings between the set sizes. However, attention paid to the 3rd digit column (i.e., the 1s column of the number sets) declined as set size increased. The difference in frequency with which participants viewed this column was not significant between sets of four ($M = 2.12$) and six ($M = 1.97$), $t(88) = 1.54, p = .13, 95\%\ CI [-.05, .36], d = .16$, but there was a significant decline in the frequency of viewing the 1s column as set size increased from six to eight ($M = 1.72$), $t(88) = 3.02, p < .01, 95\%\ CI [.08, .41], d = .33$. This decrease in focus on the 1s column from sets of six to sets of
eight corresponded to predictions for sets of eight being closer to the set means than predictions for sets of six. From these results, it appears that the combination of high attention to the 100s column and low attention to the 1s column is what caused predictions for sets of eight to be the closest to the set means. This shift in strategy (i.e., looking less at the 1s column to focus even more on the 100s column as set size grew larger) may also be the reason that response times became faster as set size increased. This may be because it seems participants traded off focusing on the more specific aspects of the sets (i.e. individuation strategy) to focus on the more general set aspects (i.e. approximation strategy) and hence took less time to process the set information, which would be similar to what Brezis et al. (2015) found with sets of sequentially presented sets of four and eight numbers. Overall, the column related shifts in strategy supported the hypothesis that participants would report that they concentrated on the 100s column of the sets.
Table 1

Means and Standard Deviations for the Strategy Description Frequency Ratings

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Sets of 4</th>
<th>Sets of 6</th>
<th>Sets of 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. “Just get a sense of the numbers.”</td>
<td>$M = 3.02$</td>
<td>$M = 3.01$</td>
<td>$M = 2.97$</td>
</tr>
<tr>
<td></td>
<td>$SD = .95$</td>
<td>$SD = .92$</td>
<td>$SD = 1.02$</td>
</tr>
<tr>
<td>2. “Find the biggest number.”</td>
<td>$M = 2.65$</td>
<td>$M = 2.63$</td>
<td>$M = 2.61$</td>
</tr>
<tr>
<td></td>
<td>$SD = 1.08$</td>
<td>$SD = 1.12$</td>
<td>$SD = 1.11$</td>
</tr>
<tr>
<td>3. “Find the smallest number.”</td>
<td>$M = 2.71$</td>
<td>$M = 2.55$</td>
<td>$M = 2.51$</td>
</tr>
<tr>
<td></td>
<td>$SD = 1.04$</td>
<td>$SD = 1.12$</td>
<td>$SD = 1.11$</td>
</tr>
<tr>
<td>4. “Try to figure out the average.”</td>
<td>$M = 2.46$</td>
<td>$M = 2.43$</td>
<td>$M = 2.38$</td>
</tr>
<tr>
<td></td>
<td>$SD = 1.11$</td>
<td>$SD = 1.20$</td>
<td>$SD = 1.15$</td>
</tr>
<tr>
<td>5. “Look for a number that is not like other numbers.”</td>
<td>$M = 1.90$</td>
<td>$M = 1.89$</td>
<td>$M = 2.02$</td>
</tr>
<tr>
<td></td>
<td>$SD = 1.00$</td>
<td>$SD = 1.02$</td>
<td>$SD = 1.03$</td>
</tr>
<tr>
<td>6. “Try to memorize specific numbers.”</td>
<td>$M = 1.46$</td>
<td>$M = 1.47$</td>
<td>$M = 1.47$</td>
</tr>
<tr>
<td></td>
<td>$SD = .85$</td>
<td>$SD = .88$</td>
<td>$SD = .92$</td>
</tr>
<tr>
<td>7. “Look at the 1st digit (e.g., the “1” in 125).”</td>
<td>$M = 3.12$</td>
<td>$M = 3.49$</td>
<td>$M = 3.49$</td>
</tr>
<tr>
<td></td>
<td>$SD = 1.03$</td>
<td>$SD = .87$</td>
<td>$SD = .94$</td>
</tr>
<tr>
<td>8. “Look at the 2nd digit (e.g., the “2” in 125).”</td>
<td>$M = 2.74$</td>
<td>$M = 2.80$</td>
<td>$M = 2.74$</td>
</tr>
<tr>
<td></td>
<td>$SD = .98$</td>
<td>$SD = 1.00$</td>
<td>$SD = .94$</td>
</tr>
<tr>
<td>9. “Look at the 3rd digit (e.g., the “5” in 125).”</td>
<td>$M = 2.12$</td>
<td>$M = 1.97$</td>
<td>$M = 1.72$</td>
</tr>
<tr>
<td></td>
<td>$SD = 1.00$</td>
<td>$SD = .98$</td>
<td>$SD = .81$</td>
</tr>
</tbody>
</table>

Note. The rating scale was as follows: 1) never, 2) some trials, 3) most trials, or 4) always.

Summary of typed descriptions. Even though there was no mention of statistics in the task instructions, 19% of the typed descriptions of processing strategy brought up the use of statistical properties (i.e., “standard deviation,” “range,” “mode,” etc.) in their
open-ended strategy descriptions. These data provided some support that adults infer the mean and variance of a set of simultaneously presented numbers when asked to estimate a prediction from that set. Certain strategies were also mentioned at least once per each set size among the descriptions collected (listed in descending order of frequency): (a) trying to average the numbers in the set (40% of descriptions), (b) looking for a pattern in the numbers (28%), (c) looking for the biggest and smallest numbers (14%), and (d) looking at the 100s column (7%). These descriptions corresponded with the five highest rated strategies from the frequency ratings (i.e., “try to figure out the average”, “just get a sense of the numbers”, “find the biggest number”, “find the smallest number”, and “look at the 1st digit”). Reports of averaging the numbers and looking at the hundreds columns supported that participants used approximation strategies (Cravalho et al., 2013; Morris et al., 2014).

About 28% of the descriptions referred to using a combination of two of the strategies listed in the previous paragraph. The most frequently cited combination was looking for the biggest and smallest numbers and trying to average the numbers. It may have been that participants who reported this combination first looked for the biggest and smallest numbers in a set and then averaged only those two numbers to make their predictions, which would represent a more efficient alternative to trying to average all the set numbers. Some participants anecdotally described employing this specific strategy. The other two reported combinations were (a) looking for patterns in the numbers and averaging the set numbers (reported once each for sets of four and sets of eight) and (b) looking at the 100s column and averaging the set numbers (reported once each for sets of
six and sets of eight). These combinations were very similar to the averaging strategies reported by Spencer’s (1963) participants, who made mean judgments of sets of simultaneously presented numbers. The fact that trying to average the set numbers was the most frequently reported strategy corroborated that adults infer the mean of a number set when asked to estimate a prediction from that set (Fowler, 1975; Pitz et al., 1976).

**Summary**

For sets of simultaneously presented numbers, predictions underestimated the set means for sets of four numbers, but overestimated the set means for sets of six and eight numbers. Set size had a significant effect on participant predictions, with sets of four differing from sets of six and eight. However, predictions for sets of six and eight did not differ. In terms of processing strategy, there was a significant increase in participant focus on the 100s column of the set numbers as set size increased from four to six, which may have contributed to the change from underestimation of predictions for sets of four to the overestimation of predictions for sets of six. There was also a significant decrease in participant focus on the 1s column of the set numbers as set size increased from six to eight, which may be why predictions for sets of eight were closer to the set means than predictions for sets of four and six, which were equal distance from the set means. Set variability and working memory had no effect on predictions and there was no interaction between set size and set variability.

Response times decreased as set size increased. Set size had a significant effect on response times, with sets of four differing from sets of six and eight, as well as sets of six differing from sets of eight. The shift in strategy from focusing less on the 1s column
to focusing more on the 100s column as set size increased may be the reason that 
response times became faster as set size increased. Sets with lower set variability elicited 
significantly faster response times than those for sets with higher set variability, but only 
for sets of eight numbers. This set variability finding also seems to be a result of 
participant strategy, as the combination of optimal strategy (i.e., participant’s highest 
focus on the 100s column and lowest focus on the 1s column) and more efficient 
encoding enabled by lower variability sets (see example in Experiment 1 hypotheses 
section) was only present for sets of eight. Finally, in terms of response times, there was 
no interaction between set size and set variability.

Addressing the research questions for Experiment 1, the strategy reports provided 
evidence that adults do infer the means and variances of number sets when estimating set 
exemplar predictions from sets of simultaneously presented numbers. Specifically, 
participants mentioned mean and variance terminology in their open-ended reports and 
indicated that they averaged set numbers and processed variance information (i.e., 
assessing set range via biggest and smallest set member, looking at the 100s column) via 
their frequency ratings and open-ended reports. However, working memory did not 
appear to have any substantial influence on adult’s predictions from sets of 
simultaneously presented numbers.

**Experiment 2**

Experiment 2 added to the literature on adult’s numerical predictions from both 
sets of simultaneously presented numbers and sets of sequentially presented numbers, as 
well as to the literature on strategies used by adults to generate numerical predictions
from sets of simultaneously presented numbers. Experiment 2 was also the first experiment to investigate adult’s strategies for generating numerical predictions from sets of sequentially presented numbers and how working memory capacity influences adult’s numerical predictions from sets of sequentially presented numbers.

**Exemplar Predictions**

Participant predictions were standardized via conversion to z-scores before being analyzed. All figures display prediction z-scores in order to clearly illustrate how far predictions were from the set means.

**Simultaneously presented numbers: Descriptive overview.** For sets of four and eight numbers participants’ predictions were, on average, within 3% of the set means, but predictions for sets of six were only within 14% of the set means (see Figure 8). Participant predictions slightly underestimated the set means for set of four numbers, but overestimated the set means for sets of six and eight numbers, which was the same general pattern seen with Experiment 1 (also simultaneously presented numbers).

When comparing predictions from sets with higher and lower set variability, the same general pattern was seen with each subset, with predictions getting further from the set means as set size increased from four to six and predictions getting closer to the set means as set size increased from six to eight. However, the between set size changes in how close predictions were to the set means were more drastic for the higher variability sets than for the lower variability sets, which again was the same general pattern seen with Experiment 1 (also simultaneously presented numbers).
Figure 8. Mean Experiment 2 prediction z-scores for sets of simultaneously presented numbers by set size and set variability. *Note.* Error bars represent SE of mean.

**Simultaneously presented numbers: Set size and set variability findings.** A 3 (set size) x 2 (set variability) repeated measures ANOVA was used to analyze the exemplar prediction data. Unlike Experiment 1, set size did not have a significant effect on how close predictions were to the set means, $F(2, 68) = 2.99, p = .06, \eta^2 = .08$. A paired-samples t-test indicated no significant difference between predictions for sets of four ($M = -.01$) and sets of eight ($M = .03$), $t < 1$. The t-test result did not support the hypothesis that predictions for sets of eight numbers would be closer to the set means than predictions for sets of four numbers, as the opposite was actually true, and therefore did not replicate the Experiment 1 results.

There was not a significant effect of set variability on exemplar predictions, $F < 1$. These findings supported the hypothesis that predictions for high and low variability sets would not differ. There also was not a significant interaction between set size and set
variability, $F(2, 68) = 2.74, p = .07, \eta^2 = .04$. All of these findings replicated the Experiment 1 results.

**Simultaneously presented numbers: Working memory findings.** To investigate if the working memory results from Experiment 1 would replicate, additional analyses were conducted. To this end, a quartile split was conducted to separate the participants into groups based on highest ($N = 8$), high ($N = 9$), low ($N = 9$), and lowest ($N = 9$) verbal WM capacity. Then a one-way ANOVA, with working memory capacity groups as a between subjects factor, was performed. WM group was not found to have an overall effect on participant predictions for sets of eight numbers, $F(3, 34) = 1.03, p = .39, \eta^2 = .10$, and Bonferroni correction states that comparisons among the four WM groups yielded no significant differences in how close predictions were to the set means. Therefore, these results replicated the Experiment 1 findings.

**Sequentially presented numbers: Descriptive overview.** Overall, predictions grew closer to the set means as set size increased (see Figure 9). Specifically, predictions for sets of four were, on average, within 13% of the set means, predictions for sets of six were within 6%, and predictions for sets of eight were within 4%. Across all set sizes, participant predictions underestimated the set means.

When comparing predictions from sets with high and low variability, the patterns were quite different. For the high variability sets, predictions got closer to the set means as set size increased from four to six and predictions got further from the set means as set size increased from six to eight. For the low variability sets, predictions also got closer to the set means as set size increased from four to six, but then predictions continue getting
closer to the set means as set size increased from six to eight. The between set size changes in how close predictions were to the means were slightly more drastic for the high variability sets than for the low variability sets.

![Figure 9](image)

*Figure 9.* Mean Experiment 2 prediction z-scores for sets of sequentially presented numbers by set size and set variability. *Note.* Error bars represent SE of mean.

**Sequentially presented numbers: Set size and set variability findings.** A 3 (set size) x 2 (set variability) repeated measures ANOVA was used to analyze the exemplar prediction data. Set size did not have a significant effect on how close predictions were to the set means, $F(2, 60) = 1.45, p = .24, \eta^2 = .04$. These results did not support the hypothesis that predictions would be closest to the set means for sets of eight numbers.

There was not a significant effect of set variability on exemplar predictions, $F < 1$. This finding did support the hypothesis that there would be no difference between
predictions for low variability and high variability sets. There also was not a significant interaction between set size and set variability, $F(2, 60) = 1.88, p = .15, \eta^2 = .06$.

**Sequentially presented numbers: Working memory findings.** To test the hypothesis that participants with more WM capacity would make predictions that are closer to the actual set means for sets of six and eight numbers, additional analyses were conducted. To this end, a quartile split was conducted to separate the participants into groups based on highest ($N = 7$), high ($N = 8$), low ($N = 8$), and lowest ($N = 8$) verbal WM capacity. Then separate one-way ANOVAs, with working memory capacity groups as a between subjects factor, were performed. WM group was not found to have an overall effect on participant predictions for sets of six numbers, $F(3, 30) = 2.67, p = .07, \eta^2 = .30$, or sets of eight numbers, $F(3, 30) = 1.52, p = .23, \eta^2 = .17$. Bonferonni correction states that comparisons among the four WM groups yielded no significant differences in how close predictions were to the set means. These results did not support the researcher’s hypothesis.

**Simultaneous v. sequential sets.** To investigate if predictions for either simultaneously (SIM) or sequentially (SEQ) presented numbers were closer to the set means, paired samples t-tests were conducted using the scores from the participants common to both the simultaneous and sequential exemplar prediction analyses ($N = 30$). Effect size corrections, for dependence between means (Morris & DeShon, 2002), were used for the paired samples t-tests. For sets of four numbers, there was no significant difference between SIM predictions ($M = -.07$) and SEQ predictions ($M = -.14$), $t(29) = 1.22, p = .23$, 95% CI [-.05, .18], $d = .22$. For sets of six numbers, there was a significant
difference between SIM predictions ($M = .12$) and SEQ predictions ($M = -.07$), $t(29) = 3.11, p < .01$, 95% CI [.06, .31], $d = .57$. Finally, for sets of eight numbers, there was no significant difference between SIM predictions ($M = .02$) and SEQ predictions ($M = -.05$), $t(29) = 1.24, p = .23$, 95% CI [-.04, .18], $d = .23$. These findings did not support the hypothesis that SIM predictions would be closer to the means than SEQ predictions, but actually provide partial support that the opposite is true.

**Response Times**

When the participant was prompted to type in how far they thought the batter would hit the next baseball, the time it took to provide a prediction was measured in milliseconds, thus all figures display participant response times in milliseconds.

**Simultaneously presented numbers: Descriptive overview.** Response times steadily decreased as set size increased (see Figure 10), which was the same general pattern seen with Experiment 1 (also simultaneously presented numbers). When comparing response times from sets with high and low variability, differing patterns were seen. For high variability sets, responses times slightly increased as set size increased from four to six numbers, but then decreased as set size increased from six to eight numbers. For low variability sets, response times decreased as set size increased.
Simultaneously presented numbers: Set size and set variability findings. A 3 (set size) x 2 (set variability) repeated measures ANOVA was used to analyze the response time data. Set size did have a significant effect on response times, $F(2, 68) = 3.60, p < .05, \eta^2 = .09$, which replicated Experiment 1. Unlike Experiment 1, within-subjects contrasts showed that there was not a significant difference between response times for sets of four and six numbers, $F(1, 34) = 1.11, p = .30, \eta^2 = .03$. However, there was a significant decrease in response times as set size increased from six to eight, $F(1, 34) = 4.58, p < .05, \eta^2 = .12$, which matched Experiment 1. Also replicating Experiment 1, a paired-samples t-test indicated a significant difference between response times for sets of four ($M = 2574.66$) and sets of eight ($M = 2368.90$), $t(34) = 2.42, p < .05$, 95% CI [33.12, 378.39], $d = .18$. 

Figure 10. Mean Experiment 2 response times for sets of simultaneously presented numbers by set size and set variability. Note. Error bars represent SE of mean.
There was not a significant effect of set variability on response times, $F < 1$. This finding did not support the hypothesis that, for sets of eight numbers, response times for low variability sets would be faster than response times for high variability sets, which also did not replicate the set variability and response time results from Experiment 1.

There also was not a significant interaction between set size and set variability, $F(2, 68) = 2.26, p = .11, \eta^2 = .04$. These results did replicate the set size and set variability interaction results from Experiment 1.

**Sequentially presented numbers: Descriptive overview.** Response times decreased as set size increased from four to six numbers, but then increased as set size increased from six to eight numbers (see Figure 11). This same general pattern was seen with sets with both high and low variability.

*Figure 11.* Mean Experiment 2 response times for sets of sequentially presented numbers by set size and set variability. *Note.* Error bars represent SE of mean.
Sequentially presented numbers: Set size and set variability findings. A 3 (set size) x 2 (set variability) repeated measures ANOVA was used to analyze the response time data. Set size did have a significant effect on response times, $F(2, 60) = 3.89, p < .05, \eta^2 = .11$. Within-subjects contrasts showed that there was a significant decrease in response times as set size increased from four to six numbers, $F(1, 30) = 8.27, p < .01, \eta^2 = .22$. However, there was no difference between response times for sets of six and eight numbers, $F < 1$. A paired-samples t-test did not indicate a significant difference between response times for sets of four ($M = 2982.09$) and sets of eight ($M = 2742.93$), $t(30) = 1.55, p = .13, 95\% \text{ CI} [-76.66, 554.99], d = .25$. There was not a significant effect of set variability on response times, $F < 1$, nor was there a significant interaction between set size and set variability, $F < 1$.

Simultaneous v. sequential sets. To investigate if response times for either simultaneously or sequentially presented numbers were faster, paired samples t-tests were conducted using the scores from the participants common to both the simultaneous and sequential exemplar prediction analyses ($N = 30$). Effect size corrections, for dependence between means (Morris & DeShon, 2002), were used for the paired samples t-tests. For sets of four numbers, there was a significant difference between SIM response times ($M = 2503.49$) and SEQ response times ($M = 2978.85$), $t(29) = -2.70, p < .05, 95\% \text{ CI} [-835.96, -114.77], d = .50$. For sets of six numbers, there was not a significant difference between SIM response times ($M = 2410.55$) and SEQ response times ($M = 2629.91$), $t(29) = -1.23, p = .23, 95\% \text{ CI} [-583.44, 144.74], d = .24$. Finally, for sets of eight numbers, there was a significant difference between SIM response times ($M = 2294.44$)
and SEQ response times ($M = 2699.04), t(29) = -3.17, p < .01, 95% CI [-665.73, -143.46], d = .60. These findings partially supported the hypothesis that SIM response times would be faster than SEQ response times.

Strategy Self-reports

Participants provided self-reports on their strategy use after each set size block of prediction task trials.

Summary of typed descriptions. Even though there was no mention of statistics in the task instructions, or the mention of figuring out an average as in the Experiment 1 frequency ratings, 23% of the typed descriptions of processing strategy brought up the use of statistical properties (i.e., “mean,” “median,” “mode,” “range,” “outlier,” etc.). These data provided support that adults infer the mean and variance of both simultaneously and sequentially presented numbers when asked to estimate a prediction from such sets. More statistical terms were brought up in these descriptions than were brought up in the Experiment 1 descriptions. Being that participants only reported one additional strategy to those from Experiment 1, it could be that more terms were reported because participants had twice as many chances to report something. It is also noteworthy that double the amount of statistical terms were brought up in strategy descriptions following simultaneous presentation (30%) compared to sequential presentation (15%) of numbers. This implies that the simultaneous presentation of numbers was more likely to provoke participants to think about the statistical properties of the sets, which may have been related to the differences in strategies for processing
sets of sequentially presented numbers and for processing sets of simultaneously presented numbers (see Table 2).

Regardless of how the set numbers were presented, trying to average the set numbers was the most frequently reported strategy, which provided further support that adults infer the means and variances of a number set when asked to estimate a prediction from that set. This finding aligned with the strategy reports from Experiment 1 and also supported that the inclusions of the description “try to figure out the average” in the Experiment 1 strategy frequency ratings didn’t influence participants to report that they averaged the set numbers for their open-ended strategy reports.

**Simultaneously presented numbers: Typed descriptions.** The most frequently described strategy was trying to average the numbers (45% of descriptions). Rounding the numbers (4%) and looking for the biggest and smallest numbers (4%) were also reported across all set sizes. Looking at the 100s column was only mentioned for sets of six and sets of eight (8%). Looking for a pattern in the numbers was mentioned for sets of four and sets of eight, but not for sets of six (4%). These strategy patterns were similar to those seen in the Experiment 1 reports (which were also based on simultaneous presentation of the numbers) in terms of the use of averaging and looking at the 100s column (see Table 2). The prominent differences were the reporting of rounding numbers, and less use of looking for patterns and looking for the biggest and smallest numbers.

About 31% of the descriptions referred to using a combination of two strategies. Four strategies were cited in combination with averaging the set numbers: (a) looking at
the 100s column (reported twice for sets of six and once for sets of eight), (b) looking for the biggest and smallest numbers (mentioned twice, but only for sets of eight), (c) rounding (mentioned twice, but only for sets of six), and (d) memorizing numbers (reported once for sets of four). From this list, only the combination of memorizing and averaging numbers was not reported by any of the Experiment 1 (also simultaneous presentation) participants. However, this strategy appeared to be idiosyncratic, given only one participant across two studies reported using it.

**Sequentially presented numbers: Typed descriptions.** The most frequently described strategy was also trying to average the numbers (41% of descriptions). Looking at the 100s column (14%), looking for a pattern in the numbers (11%), and rounding numbers (4%) were also mentioned across all set sizes. About 18% of the descriptions referred to using a combination of two strategies. Four strategies were cited in combination with averaging the set numbers: (a) rounding (reported once each for sets of four and sets of eight), (b) looking at the 100s column (reported once for sets of four), (c) memorizing numbers (reported once for sets of six), and (d) ascertaining the set range (mentioned once for sets of six). Although reported with less frequency, this list is comprised of basically the same strategy combinations for processing sets of sequentially presented numbers to those reported by participants after processing sets of simultaneously presented numbers. The only obvious differences are that looking for the biggest and smallest numbers was reported for processing sets of simultaneously presented numbers and ascertaining the set range was reported for processing sets of sequentially presented numbers. However, one could argue these are two ways to phrase
the same strategy being that the biggest and smallest set numbers are what give you the set range.

**Simultaneous v. sequential sets.** Supporting the hypothesis that strategies would differ based on presentation type, there were prominent differences in strategy between making predictions for simultaneously and sequentially presented numbers (see Table 2). First, sequential presentation prompted more looking for patterns in the numbers and looking at the 100s column. However, it seems that simultaneous presentation prompts more focus on the 100s column for larger sets (i.e., six and eight numbers). For Experiment 1 (simultaneous presentation only), participants reported paying more attention to the 100s column for sets of six and eight numbers, and when the numbers were presented simultaneously in Experiment 2 participants only reported looking at the 100s for sets of six and eight. However, when the numbers were presented sequentially in Experiment 2, participants reported looking at the 100s for all set sizes. Second, only simultaneous presentation prompted looking for the biggest and smallest numbers in the set. Looking for the biggest and smallest numbers is a strategy that implies one was trying to ascertain more diagnostic variance information (i.e., the exact set range) about the set. Although no participant described such a strategy, it could be that when the numbers were presented simultaneously, some participants first looked at the 100s column to seek general variance information and then sought more specific variance information by looking for the set range via the biggest and smallest numbers in the set. This more statistically precise processing strategy may explain why twice the amount of
statistical terms appeared in strategy reports from the simultaneous presentation of the set numbers as compared to reports from the sequential presentation of set numbers.

In addition, the rounding of set numbers seemed to be prompted by the sequential presentation of the numbers. Rounding was not reported as a strategy by any Experiment 1 (simultaneous presentation only) participants, but was reported by the same amount of participants from both simultaneous and sequential presentation strategy descriptions in Experiment 2. It seems that the same participants first used rounding when presented with one number at a time and then applied that strategy to when the numbers were presented all at once. Therefore, it is possible that no one in Experiment 1 reported rounding because they did not think to use that strategy without having to process numbers sequentially first. Finally, it was thought that when the numbers were presented sequentially, participants may try to round set numbers to the 100s column to make it easier to hold previously seen set numbers in working memory as they updated their prediction based on the number they are currently encoding. However, participants did not report paying more attention to the 10s column (i.e., in order to round to the 100s column) when the numbers were presented sequentially. In fact, the adults did not report any of the specifics of how they rounded the numbers.
Table 2

Proportion of Participants Descriptions that described Listed Strategies

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Experiment 1: Simultaneous Presentation</th>
<th>Experiment 2: Simultaneous Presentation</th>
<th>Experiment 2: Sequential Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaging set numbers.</td>
<td>40%</td>
<td>45%</td>
<td>41%</td>
</tr>
<tr>
<td>Looking for a pattern in the set numbers.</td>
<td>28%</td>
<td>4%</td>
<td>11%</td>
</tr>
<tr>
<td>Looking for the biggest and smallest set numbers.</td>
<td>14%</td>
<td>4%</td>
<td>0%</td>
</tr>
<tr>
<td>Looking at the 100s column.</td>
<td>7%</td>
<td>8%</td>
<td>14%</td>
</tr>
<tr>
<td>Rounding numbers.</td>
<td>0%</td>
<td>4%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Summary for Simultaneously Presented Numbers

Predictions underestimated the set means for sets of four numbers, but overestimated the set means for sets of six and eight numbers, which matched the Experiment 1 results. However, unlike Experiment 1, set size did not have a significant effect on participant predictions. Also replicating Experiment 1, set variability and working memory had no effect on predictions and there was no interaction between set size and set variability.

Response times decreased as set size increased, which matched the Experiment 1 findings. Set size had a significant effect on response times, with sets of eight differing from sets of four and sets of six. However, unlike Experiment 1, response times for sets
of four did not differ from those for sets of six. Also unlike Experiment 1, set variability had no effect on response times. In addition, there was no interaction between set size and set variability, which replicated Experiment 1.

It may be that the Experiment 2 simultaneous presentation results did not fully replicate the Experiment 1 (simultaneously presentation only) results because the larger sample of Experiment 1 participants skewed those results. Much more than necessary statistical power might explain why significant effects of set size on predictions and set variability on response times were found for Experiment 1, but not Experiment 2. It is likely that the Experiment 2 results, which were based on the exact recommended amount of participants from a power analysis, may be more indicative of predictions from simultaneously presented numbers.

**Summary for Sequentially Presented Numbers**

Predictions underestimated the set means for all set sizes. Neither set size, nor set variability, or working memory had a significant effect on participant predictions, and there was no interaction between set size and set variability.

Response times decreased as set size increased from four to six numbers, but increased as set size increased from six to eight numbers. Set size had a significant effect on response times, with sets of four differing from sets of six. However, response times for sets of eight did not differ from those for sets of four or six. Set variability had no effect on predictions, and there was no interaction between set size and set variability.
Simultaneous v. Sequential Sets

For sets of six numbers, SEQ predictions were significantly closer to the set means than SIM predictions. However, SIM response times were faster than SEQ response times, but only for sets of four and eight numbers.

Strategy Summary

Addressing the first research questions for Experiment 2, the strategy reports provided evidence that adults do infer the means and variances of number sets when estimating set exemplar predictions from sets of sequentially presented numbers. These reports also provided more evidence that adults infer the means and variances from sets of simultaneously presented numbers. Specifically, participants mentioned mean and variance terminology and indicated that they averaged set numbers and processed variance information (i.e., looked at the 100s column) via their open-ended reports for both simultaneously and sequentially presented numbers.

Addressing the second and third research questions for Experiment 2, working memory did not appear to have any substantial influence on adult’s predictions from sets of sequentially presented numbers, but the simultaneous and sequential presentation of set numbers had differing influences on adult prediction strategies. Simultaneous presentation prompted more focus on the 100s column for larger sets, as participants reported looking at the 100s only for sets of six and eights, but reported looking at the 100s for all set sizes when the numbers were presented sequentially. Also, only simultaneous presentation of set numbers prompted looking for the biggest and smallest
numbers in the set, and sequential presentation of set numbers prompted the rounding of set numbers.

Experiment 3

Experiment 3 added to the literature on adult’s numerical predictions from both sets of simultaneously presented numbers and sets of sequentially presented numbers. Experiment 3 was also the first experiment to investigate children’s numerical predictions from both sets of simultaneously presented numbers and sets of sequentially presented numbers, as well as children’s strategies for generating numerical predictions from both sets of simultaneously presented numbers and sets of sequentially presented numbers. Finally, Experiment 3 was also the first experiment to investigate how working memory capacity influences children’s numerical predictions from both sets of simultaneously presented numbers and sets of sequentially presented numbers.

Problematic Eye-tracking Data

It was intended to triangulate the Experiment 3 adult eye-tracking data with the adult self-reported strategies from Experiments 1 and 2. Another objective was to compare the adult and child eye-tracking data to see the differences in their attention to specific number set information during encoding. However, in preparing the adult and child eye-tracking data for coding and analysis, two problems arose that prevented triangulation of the eye-tracking and self-report data, as well as the comparison of the adult and child eye-tracking data.

One problem was that poor overall acuity for the recordings of participant eye fixations during the sequential version of the exemplar prediction task left few recordings
that could be objectively coded. Specifically, the minimum acuity level was set at 50%, being that, when using a Tobii® T-60XL eye tracker, this is the level at which the fixations of at least one eye is tracked for the whole recording. For the recordings of participant eye fixations during the sequential version of the exemplar prediction task, only five adult ($M = 64\%$) and two child ($M = 62\%$) recordings exceeded this acuity criterion. Consequently, too little eye-tracking data from the sequential version of the exemplar prediction task was left to justify inclusion in the Experiment 3 analysis.

There were no acuity problems with the recordings of participant eye fixations during the simultaneous version of the exemplar prediction task. Recordings from 27 adults ($M = 84\%$) and 25 children ($M = 75\%$) exceeded the acuity criterion. However, a different problem also left too little of this data to justify inclusion in the Experiment 3 analysis. Namely, an uncorrectable error was detected with most of the recordings of participant eye fixations during both versions of the exemplar prediction task. After double-checking each recording, it was found that 75% of them were out of sync and could not be objectively coded. Multiple alternatives to coding the data from the recordings were investigated, but neither was found to alleviate the syncing problem and provide an objective way to code these data. Therefore, no analysis of the eye-tracking data will be discussed further in this document.

**Possible Task Order Effects**

Another problem arose when implementing exemplar prediction tasks via the Tobii® T-60XL eye tracking monitor and software. After running the sequential presentation version of the prediction task, it would take 10 minutes or longer for the eye.
tracking data to save. Being that the simultaneous presentation and sequential presentation versions of the number set prediction task were counterbalanced, if the participant had completed the sequential version before the simultaneous version, the participant would have to wait extra time before completing the simultaneous version. With the first few participants this was not much of a problem, but the length of the data saving delay increased as more eye fixation recordings were collected. This forced the experimenter to run the delay free simultaneous version of the task first with more of the remaining participants so that they could finish both tasks in the allotted time.

For this reason, the exemplar task version counterbalance became uneven for the adults and children participating in Experiment 3. For adults, 26 participants completed the simultaneous presentation version first, with only 11 participants completing the sequential presentation version first. For children, 23 participants completed the simultaneous presentation version first, with 12 participants completing the sequential presentation version first. The effect of this task order imbalance was investigated and will be discussed below.

**Number Line Estimations**

For the 0-1,000 number line estimation task, percent absolute error (PAE) was calculated to measure estimation accuracy (i.e., how close a hatch mark was to where the number would actually fall on the line). The average PAE was 4.04% (SD = 2.33%) for the adult estimations and 6.30% (SD = 3.69%) for the child estimations. Linear and logarithmic functions were also fit to each set of participant estimations. It was the case for all of the adult estimations that the best-fitting linear function was a better fit than the
best-fitting logarithmic function. Therefore, it was no surprise that the aggregated adult estimations fit a linear function better than a logarithmic one ($R^2 = 1.00$ vs. $R^2 = .65$; see Figure 12). All of the child estimations, except for one, fit a linear function better than a logarithmic one. The child’s data that was a better fit to a logarithmic function also turned out to be an outlier based on their exemplar prediction data, so all of their data was removed from the further analysis. The remaining aggregated child estimations fit a linear function better than a logarithmic one ($R^2 = .99$ vs. $R^2 = .68$; see Figure 13).

Figure 12. Aggregated adult number line estimations.
Figure 13. Aggregated child number line estimations.

**Adult Exemplar Predictions**

Participant predictions were standardized via conversion to z-scores before being analyzed. All figures display prediction z-scores in order to clearly illustrate how far predictions were from the set means.

**Simultaneously presented numbers: Descriptive overview.** Predictions for sets of eight were, on average, within 5% of the sets means, whereas those for sets of four were within 9% of the set means (see Figure 14). Predictions for sets of six were only within 14% of the set means, which exactly replicated the results of Experiment 2. Participant predictions underestimated the set means for set of four numbers, but overestimated the set means for sets of six, which was the same general pattern seen with Experiment 2 simultaneous presentation results. However, unlike the Experiment 2 results, predictions for set of eight underestimated the set means, but this may have been due to the task order imbalance among participants in Experiment 3.
When comparing predictions from sets with high and low set variability, the patterns are quite different. Predictions for sets with high variability got further from the set means as set size increased from four to six, but then predictions got closer to the set means as set size increased from six to eight. Predictions for the low variability sets were the same distance from the mean as set size grew from four to six, but then predictions got closer to the set means as set size increased from six to eight. The between set size changes in how close predictions were to the set means were also more drastic for the high variability sets than for the low variability sets, which was the same general pattern seen in the Experiment 2 simultaneous presentation results.

![Figure 14](image)

*Figure 14.* Mean Experiment 3 adult prediction z-scores for sets of simultaneously presented numbers by set size and set variability. *Note.* Error bars represent SE of mean.

**Simultaneously presented numbers: Set size and set variability findings.** A 3 (set size) x 2 (set variability) repeated measures ANOVA was used to analyze the simultaneous exemplar prediction data. Unlike Experiment 2, set size did have a
significant effect on how close predictions were to the set means, \( F(2, 70) = 9.77, p < .001, \eta^2 = .22 \). Within-subjects contrasts showed that predictions for sets of four and sets of six were significantly different, \( F(1, 35) = 13.71, p < .01, \eta^2 = .28 \). This difference reflects that predictions for sets of four were closer to the set means, but also the contrast between the underestimation of sets of four and the overestimation of sets of six. Within-subject contrasts also showed a significant difference between predictions for sets of six and sets of eight, \( F(1, 35) = 10.66, p < .01, \eta^2 = .23 \). A paired-samples t-test indicated no significant difference between predictions for sets of four (\( M = -.09 \)) and sets of eight (\( M = -.05 \)), \( t < 1 \), which replicated Experiment 2.

There was not a significant effect of set variability on exemplar predictions, \( F(1, 35) = 2.09, p = .16, \eta^2 = .06 \). These findings supported the hypothesis that predictions for high and low variability sets would not differ. There also was not a significant interaction between set size and set variability, \( F < 1 \). All of these findings replicated the Experiment 2 results.

**Simultaneously presented numbers: Task order findings.** Independent samples t-tests were conducted to test for an effect of the task order imbalance discussed above. Namely, more participants (\( N = 25 \)) completed the SIM version of the prediction task first than those who completed the SEQ version first (\( N = 11 \)). For sets of four numbers, no significant difference was found between the SIM first group (\( M = -.10 \)) and SEQ first group (\( M = -.05 \)), \( t < 1 \). For sets of six numbers, no significant difference was found between the SIM first group (\( M = .12 \)) and SEQ first group (\( M = .19 \)), \( t < 1 \). Finally, for sets of eight numbers, no significant difference was found between the SIM
first group \((M = -0.05)\) and SEQ first group \((M = -0.06)\), \(t < 1\). To summarize, task order had no effect on predictions made from sets of simultaneously presented numbers.

**Sequentially presented numbers: Descriptive overview.** Predictions became much closer to the set means as set size increased from four (within 16%) to six numbers (within 1%), then stayed very close to the set means for sets of eight (within 2%; see Figure 15). Across all set sizes, participant predictions underestimated the set means. These were the same general patterns seen for predictions from sequentially presented numbers from Experiment 2.

When comparing predictions from sets with high and low variability, the patterns were quite different. For the high variability sets, predictions got much closer to the set means as set size increased from four to six and then moderately closer to the set means as set size increased from six to eight. For the low variability sets, predictions got further from the set means as set size increased from four to six, but then predictions got slightly closer to the set means as set size increased from six to eight. The between set size changes in how close predictions were to the set means were more drastic for the high variability sets than for the low variability sets. These set variability based patterns were quite different from those seen for sequentially presented numbers from Experiment 2, but this may have been due to the task order imbalance among participants in Experiment 3.
Sequentially presented numbers: Set size and set variability findings. A 3 (set size) x 2 (set variability) repeated measures ANOVA was used to analyze the sequential exemplar prediction data. Unlike Experiment 2, there was a significant interaction between set size and set variability, \( F(2, 70) = 10.99, p < .001, \eta^2 = .23 \). Within-subjects contrasts showed that as set size increased from four to six, predictions for high variability sets became significantly closer to the set means as predictions for low variability sets became significantly further from the set means, \( F(1, 35) = 21.13, p < .001, \eta^2 = .38 \). However, as set size increased from six to eight, there were no significant changes in how close predictions for high or low variability sets were to the set means, \( F < 1 \).

Also unlike Experiment 2, set size did have a significant effect on how close predictions were to the set means, \( F(2, 70) = 9.59, p < .001, \eta^2 = .17 \). Within-subjects
contrasts showed that predictions for sets of four and sets of six were significantly
different, $F(1, 35) = 13.29, p < .01, \eta^2 = .28$. However, within-subject contrasts showed
no difference between predictions for sets of six and sets of eight, $F < 1$. A paired-
samples t-test also indicated a significant difference between predictions for sets of four
($M = -.16$) and sets of eight ($M = -.02$), $t(35) = -3.92, p < .001, 95\% \text{ CI } [-.21, -.07], d = .66$, which was again unlike Experiment 2. These results did not support the hypothesis
that predictions would not differ based on set size. Finally, there was not a significant
effect of set variability on exemplar predictions, $F(1, 35) = 1.77, p = .19, \eta^2 = .03$. This
finding supported the hypothesis that there would be no difference between predictions
for low variability and high variability sets.

**Sequently presented numbers: Task order findings.** Independent samples t-
tests were conducted to test for an effect of the task order imbalance discussed above.
Namely, more participants ($N = 26$) completed the SIM version of the prediction task first
than those who complete the SEQ version first ($N = 10$). For sets of four numbers, no
significant difference was found between the SIM first group ($M = -.14$) and SEQ first
group ($M = -.22$), $t(34) = 1.03, p = .31, 95\% \text{ CI } [-.08, .23], d = .39$. For sets of six
numbers, no significant difference was found between the SIM first group ($M = .02$) and
SEQ first group ($M = -.08$), $t < 1$. Finally, for sets of eight numbers, no significant
difference was found between the SIM first group ($M = -.02$) and SEQ first group ($M = -.02$), $t < 1$. To summarize, task order had no effect on predictions made from sets of
sequentially presented numbers.
Simultaneous v. sequential sets. To investigate if predictions for either simultaneously or sequentially presented numbers were closer to the mean, paired samples t-tests were conducted using the scores from the participants common to both the simultaneous and sequential exemplar prediction analyses ($N = 34$). Effect size corrections, for dependence between means (Morris & DeShon, 2002), were used for the paired samples t-tests. For sets of four numbers, there was no significant difference between SIM predictions ($M = -.07$) and SEQ predictions ($M = -.15$), $t(33) = 1.82, p = .08$, 95% CI [-.01, .17], $d = .31$. For sets of six numbers, there was no significant difference between SIM predictions ($M = .11$) and SEQ predictions ($M = .01$), $t(33) = 1.74, p = .09$, 95% CI [-.02, .23], $d = .30$. This finding did not support the hypothesis that SEQ predictions from sets of six numbers would be closer to the set means than SIM prediction from sets of six numbers. Finally, for sets of eight numbers, there was no significant difference between SIM predictions ($M = -.05$) and SEQ predictions ($M = .00$), $t(33) = -1.11, p = .28$, 95% CI [-.13, .04], $d = .20$.

Child Exemplar Predictions

Participant predictions were standardized via conversion to z-scores before being analyzed. All figures display prediction z-scores in order to clearly illustrate how far predictions were from the set means.

Simultaneously presented numbers: Descriptive overview. Predictions for sets of eight were, on average, within 18% of the sets means, whereas those for sets of four were within 21% of the set means (see Figure 16). Predictions for sets of six were only within 38% of the set means. Although children’s predictions were, per each set size, at
least 12% further away from the set means as the Experiment 3 adult’s predictions, the
general set size pattern matched the adult’s results. Children’s predictions
underestimated the set means for set of four numbers, but overestimated the set means for
sets of six and eight numbers, which matched the general pattern seen with the adult
Experiment 1 and 2 simultaneous presentation results.

When comparing predictions from sets with high and low variability, the patterns
are quite different. Predictions for sets with high variability got further from the set
means as set size increased from four to six, but then predictions got closer to the set
means as set size increased from six to eight. Predictions for the low variability sets got
closer to the set means as set size increased from four to six, but then predictions got
further from the set means as set size increased from six to eight. The between set size
changes in how close predictions were to the means were even more drastic for both high
variability and low variability sets than the changes in adult predictions.
Simultaneously presented numbers: Set size and set variability findings. A 3 (set size) x 2 (set variability) repeated measures ANOVA was used to analyze the simultaneous exemplar prediction data. Unlike the adult results, there was a significant interaction between set size and set variability, $F(2, 68) = 7.32, p < .05, \eta^2 = .10$. Within-subjects contrasts showed that as set size increased from four to six, predictions for high variability sets became significantly further from the set means as predictions for low variability sets became significantly closer to the set means, $F(1, 34) = 7.84, p < .05, \eta^2 = .19$. However, as set size increased from six to eight, there were no significant changes in how close predictions for high or low variability sets were to the set means, $F(1, 34) = 2.76, p = .11, \eta^2 = .08$.

As it did for adult’s predictions, set size did have a significant effect on how close children’s predictions were to the set means, $F(2, 68) = 4.14, p < .05, \eta^2 = .10$. Within-
subjects contrasts showed that predictions for sets of four and sets of six were significantly different, $F(1, 34) = 6.59, p < .05, \eta^2 = .16$. However, within-subject contrasts also showed no significant difference between predictions for sets of six and sets of eight, $F(1, 34) = 1.39, p = .25, \eta^2 = .04$, which was unlike the adult findings. A paired-samples t-test indicated no significant difference between predictions for sets of four ($M = -.21$) and sets of eight ($M = .18$), $t(34) = -1.77, p = .09, 95\% \text{ CI} [-.83, .06], d = .31$, which matched the adult results. There was not a significant effect of set variability on exemplar predictions, $F(1, 34) = 1.34, p = .26, \eta^2 = .03$, which mirrored the adult findings.

**Simultaneously presented numbers: Working memory findings.** To investigate if children with higher working memory capacity make predictions that are closer to the actual set means than those with lower capacity, additional analyses were conducted. To this end, a quartile split was conducted to separate the participants into groups based on highest ($N = 6$), high ($N = 2$), low ($N = 13$), and lowest ($N = 7$) verbal WM capacity. Then separate one-way ANOVAs, with working memory capacity groups as a between subjects factor, were performed. WM group was not found to have an overall effect on participant predictions for sets of four, $F(3, 27) = 2.11, p = .13, \eta^2 = .26$, sets of six, $F(3, 27) = 2.16, p = .12, \eta^2 = .27$, or sets of eight numbers, $F(3, 27) = 2.10, p = .13, \eta^2 = .26$. Bonferroni correction states that comparisons among the four WM groups yielded no significant differences in how close predictions were to the set means. Therefore, the hypothesis that children with higher working memory capacity would
make predictions closer to the set means than those with lower capacity was not supported.

**Simultaneously presented numbers: Task order findings.** Independent samples t-tests were conducted to test for an effect of the task order imbalance discussed above. Namely, more participants ($N = 23$) completed the SIM version of the prediction task first than those who complete the SEQ version first ($N = 12$). For sets of four numbers, no significant difference was found between the SIM first group ($M = -.42$) and SEQ first group ($M = .19$), $t < 1$. For sets of six numbers, no significant difference was found between the SIM first group ($M = .37$) and SEQ first group ($M = .42$), $t < 1$. Finally, for sets of eight numbers, no significant difference was found between the SIM first group ($M = .04$) and SEQ first group ($M = .44$), $t < 1$. To summarize, task order had no effect on predictions made from sets of simultaneously presented numbers.

**Sequentially presented numbers: Descriptive overview.** Predictions became closer to the set means as set size increased from four (within 11%) to six numbers (within 9%), but predictions for sets of eight were much further from the set means (within 45%; see Figure 17). Children’s predictions were actually closer to the set means for sets of four than the Experiment 3 adult's predictions, who were within 16%. However, children’s predictions were 8% further from the set means than adult’s predictions for sets of six and a noteworthy 43% further from the set means than adult’s predictions for sets of eight. It is difficult to tell why children’s predictions for sets of eight were so much further from the mean than adult’s predictions for sets of eight. One explanation is that processing sets of eight numbers is more difficult for children than for
adults due to children having more limited working memory capacity. Although it may be that fatigue also played a role in children’s poor performance because sets of eight were always presented last.

Predictions for sets of four and sets of six underestimated the set means, but predictions for sets of eight overestimated the set means. These patterns for children’s predictions from sequentially presented numbers were quite different from the adult Experiment 2 and 3 sequential presentation results.

When comparing predictions from sets with high and low variability, the patterns were direct opposites. For the high variability sets, predictions got much closer to the set means as set size increased from four to six and then moderately closer to the set means as set size increased from six to eight. For the low variability sets, predictions got much further from the set means as set size increased from four to six and then exceedingly further from the set means as set size increased from six to eight. The between set size changes in how close predictions were to the set means were more drastic for the low variability sets than for the high variability sets. Besides the pattern for high variability sets, these set variability based patterns were quite different from the adult Experiment 2 and 3 sequential presentation results.
Figure 17. Mean Experiment 3 child prediction z-scores for sets of sequentially presented numbers by set size and set variability. Note. Error bars represent SE of mean.

Sequentially presented numbers: Set size and set variability findings. A 3 (set size) x 2 (set variability) repeated measures ANOVA was used to analyze the sequential exemplar prediction data. As it did for adult’s predictions, there was a significant interaction between set size and set variability, $F(2, 68) = 6.47, p < .01, \eta^2 = .09$. Within-subjects contrasts showed that as set size increased from four to six, there were no significant changes in how close predictions for high or low variability sets were to the set means, $F(1, 34) = 3.62, p = .07, \eta^2 = .10$. However, as set size increased from six to eight, predictions for high variability sets became significantly closer to the set means as predictions for low variability sets became significantly further from the set means, $F(1, 34) = 9.27, p < .01, \eta^2 = .21$. These within-subject contrasts findings were the opposite of the adult findings.
Also like adult’s predictions, set size did have a significant effect on how close children’s predictions were to the set means, $F(2, 68) = 4.29, p < .05, \eta^2 = .10$. Within-subjects contrasts showed that predictions for sets of four and sets of six were not significantly different, $F < 1$, which was unlike the adult results. Within-subject contrasts also showed a significant difference between predictions for sets of six and sets of eight, $F(2, 68) = 5.39, p < .05, \eta^2 = .14$, which was also unlike the adult results. A paired-samples t-test also indicated a significant difference between predictions for sets of four ($M = -.11$) and sets of eight ($M = .45$), $t(34) = -2.60, p < .05, 95\% \text{ CI} [-1.00, -.12], d = .44$, which matched the adult results. There was not a significant effect of set variability on exemplar predictions, $F < 1$, which mirrored the adult findings.

**Sequentially presented numbers: Working memory findings.** To investigate if children with higher working memory capacity make predictions that were closer to the actual set means than those with lower capacity, additional analyses were conducted. To this end, a quartile split was conducted to separate the participants into groups based on highest ($N = 6$), high ($N = 2$), low ($N = 13$), and lowest ($N = 7$) verbal WM capacity. Then separate one-way ANOVAs, with working memory capacity groups as a between subjects factor, were performed. WM group was not found to have an overall effect on participant predictions for sets of four, $F(3, 27) = 1.36, p = .26, \eta^2 = .17$, sets of six, $F(3, 27) = 2.24, p = .11, \eta^2 = .28$, or sets of eight numbers, $F(3, 27) = 2.67, p = .07, \eta^2 = .33$. Bonferroni correction states that comparisons among the four WM groups yielded no significant differences in how close predictions were to the set means. Therefore, the
hypothesis that children with higher working memory capacity would make predictions
closer to the set means than those with lower capacity was not supported.

**Sequentially presented numbers: Task order findings.** Independent samples t-
tests were conducted to test for an effect of the task order imbalance discussed above.
Namely, more participants \((N = 23)\) completed the SIM version of the prediction task first
than those who complete the SEQ version first \((N = 12)\). For sets of four numbers, no
significant difference was found between the SIM first group \((M = -.18)\) and SEQ first
group \((M = -.03)\), \(t < 1\). For sets of six numbers, no significant difference was found
between the SIM first group \((M = -.02)\) and SEQ first group \((M = -.21)\), \(t < 1\). Finally, for
sets of eight numbers, no significant difference was found between the SIM first group
\((M = .35)\) and SEQ first group \((M = .66)\), \(t < 1\). To summarize, task order had no effect
on predictions made from sets of sequentially presented numbers.

**Simultaneous v. sequential sets.** To investigate if predictions for either
simultaneously or sequentially presented numbers were closer to the mean, paired
samples t-tests were conducted using the scores from the participants common to both the
simultaneous and sequential exemplar prediction analyses \((N = 35)\). Effect size
corrections, for dependence between means (Morris & DeShon, 2002), were used for the
paired samples t-tests. For sets of four numbers, there was no significant difference
between SIM predictions \((M = -.21)\) and SEQ predictions \((M = -.11)\), \(t < 1\). For sets of
six numbers, there was no significant difference between SIM predictions \((M = .38)\) and
SEQ predictions \((M = -.09)\), \(t(34) = 1.92, p = .06, 95\% CI [-.03, .97], d = .33\). Finally, for
sets of eight numbers, there was no significant difference between SIM predictions ($M = .18$) and SEQ predictions ($M = .45$), $t(34) = -1.06$, $p = .30$, 95% CI [-.80, .25], $d = .18$.

**Adult v. Child Predictions**

To test the hypothesis that child predictions would be further from the set means than adult predictions, independent samples $t$-tests were conducted for both sets of simultaneously and sequentially presented numbers.

**Simultaneously presented numbers.** For sets of four numbers, there was no significant difference between adult predictions ($M = -.09$) and child predictions ($M = -.21$), $t < 1$. For sets of six numbers, there was no significant difference between adult predictions ($M = .14$) and child predictions ($M = .38$), $t < 1$. For sets of eight numbers, there was no significant difference between adult predictions ($M = -.05$) and child predictions ($M = .18$), $t < 1$. These results did not support the experimenter’s hypothesis.

**Sequentially presented numbers.** For sets of four numbers, there was no significant difference between adult predictions ($M = -.16$) and child predictions ($M = -.11$), $t < 1$. For sets of six numbers, there was no significant difference between adult predictions ($M = -.01$) and child predictions ($M = -.09$), $t < 1$. For sets of eight numbers, there was no significant difference between adult predictions ($M = -.02$) and child predictions ($M = .45$), $t < 1$. These results did not support the experimenter’s hypothesis.

**Child Strategy Self-reports**

Participants provided self-reports on their strategy use after completing the entire prediction task.
**Summary of typed descriptions.** Strategy self-reports were not elicited from the first 10 child participants. Therefore, only 75% of the participants verbally reported on the strategies they used to complete each version of the exemplar prediction task. Unlike the adult strategy self-reports from Experiments 1 and 2, where around 20% of participants specifically mentioned descriptive statistics (i.e., “mean,” “median,” “mode,”), none of the child participants mentioned descriptive statistics. However, regardless of how the set numbers were presented, the children did report strategies indicating they paid attention to set variance (i.e., looking at the 100s, rounding numbers, looking for the biggest and smallest numbers), which provides some support that they inferred the variances of the number sets.

Another major difference was how precise adults and children were in describing their averaging strategies. Only one child participant used the exact phrasing that they “tried to average” the set numbers in describing their strategy, whereas 26% of the adults in Experiment 1 and approximately 46% of the adults in Experiment 2 used specific phrasing about trying to average the set numbers (regardless of set size or presentation type). However, 35% of the children used other phrasing that implied they were trying to average the numbers (e.g., they described trying to find a number “in between” or “in the middle of” the set numbers), providing some evidence that they were inferring the means of the number sets. Few adults (three for Experiment 1; between three and eight for Experiment 2 depending on set size and presentation type) used such alternative phrasing to describe trying to average the numbers.
**Simultaneously presented numbers: Typed descriptions.** The most frequently described strategy was looking for a pattern in the numbers (43% of descriptions; see Table 3). However, the only pattern that was specified was looking for if the numbers were getting “bigger or smaller” (17%). The next most frequently reported strategies were to look at the 100s column and then the 10s column (22%). Other strategies that were reported include phrasing that implies averaging the numbers (17%), looking at the 100s column (17%), scanning the 100s, 10s, and 1s columns (17%), and looking for the biggest and smallest numbers (9%). Another common strategy was rounding the numbers (17%), with some participants specifying that they either rounded from the 10s to the 100s (15%) or from the 1s to the 10s (11%). Although the strategies listed above all seem to be related to the task goal of making a prediction via range and variance information, some other strategies were reported that seem at odds with the task goal. For instance, children reported looking only at the last few numbers (22%), as well as using addition (9%), subtraction (4%), or counting (4%).

Only four descriptions referred to using a combination of two strategies. The following three strategies were cited in combination with averaging the set numbers: (a) looking for the biggest and smallest numbers (mentioned twice), (b) rounding, and (c) looking at the 100s column. This list is comprised of strategy combinations also reported by adults for processing all the set numbers at once.

**Sequentially presented numbers: Typed descriptions.** The most frequently described strategy was looking at the 100s column (39% of descriptions; see Table 3). The next most frequently reported strategy was looking for a pattern in the numbers
with the only pattern that was specified again being looking for the if the numbers were getting “bigger or smaller” (17%). Many participants also reported scanning the 100s, 10s, and 1s columns (35%) or implied that they were trying to average the numbers (30%). Although reported less frequently, other reported strategies overlapped with those used when the numbers were presented simultaneously, namely to look at the 100s column and then the 10s column (13%) and looking for the biggest and smallest numbers (4%). The use of rounding the numbers (9%) was also reported less frequently, again with some participants specifying that they either rounded from the 10s to the 100s (4%) or from the 1s to the 10s (4%). Another difference was that participants reported using strategies seemingly at odds with the task goal more often, specifically, using addition (13%), subtraction (13%), or counting (9%). However, the use of looking only at the last few numbers (17%) was reported less frequently. Finally, only one unique strategy emerged from the sequential version self-reports, which was trying to memorize numbers (9%).

It may be that the strategy reports describing the use of addition (for both sets of simultaneously and sequentially presented numbers) were all reported by a specific segment of child participants. Specifically, a small group of children misunderstood the goal of the sequential version of the exemplar prediction task due to a specific design element. Specifically, a fixation cross was shown before each set number as each of the set of numbers was presented (see Figure 4). Two points of evidence suggested that these children interpreted the fixations crosses as plus signs and then tried to add up the sequentially presented numbers. These children would take longer to verbalize their
predictions, suggesting slower addition strategies over faster estimation strategies (Siegler, 1996), and their predictions consisted of four digits or more, whereas predictions from other children were almost always three digits. To test if the children suspected of adding up the numbers were indeed doing so, the differences between these children’s predictions and both the set means and the sum of the set numbers were calculated. A paired samples t-test, with effect size corrections for dependence between means (Morris & DeShon, 2002), confirmed that the predictions of the children suspected of adding up the numbers were significantly closer to the set sums ($M = 1491.04$) than the set means ($M = 2543.13$), $t(23) = 2.58$, $p < .05$, $d = .54$.

Only four descriptions referred to using a combination of two strategies. The following three strategies were cited in combination with averaging the set numbers: (a) looking for the biggest and smallest numbers (mentioned twice), (b) looking at the 100s column, and (c) memorizing numbers. This list is comprised of strategy combinations also reported by adults for processing sets of sequentially presented numbers.
### Table 3

*Proportion of Participants Descriptions that described Listed Strategies*

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Experiment 2 (Adult self-reports)</th>
<th>Experiment 3 (Child self-reports)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averaging set numbers.</td>
<td>45%</td>
<td>41%</td>
</tr>
<tr>
<td>Looking at the 100s column.</td>
<td>8%</td>
<td>14%</td>
</tr>
<tr>
<td>Looking for a pattern in the set numbers.</td>
<td>4%</td>
<td>11%</td>
</tr>
<tr>
<td>Rounding numbers.</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>Looking for the biggest and smallest set numbers.</td>
<td>4%</td>
<td>0%</td>
</tr>
<tr>
<td>Looking at the 100s, 10s and 1s columns.</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Looking at the last few numbers in the set.</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Looking at the 100s column, then looking at the 10s column.</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Adding numbers.</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Subtracting numbers.</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Counting.</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Memorizing numbers.</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Summary for Adult Predictions

When the numbers were presented simultaneously, predictions underestimated the set means for sets of four numbers and overestimated the set means for sets of six numbers, which matched the Experiment 2 results. However, unlike Experiment 2, predictions underestimated the set means for sets of eight numbers. Also unlike Experiment 2, set size did have a significant effect on participant predictions, with predictions for sets of four significantly differing from those for sets of six and sets of six significantly differing from those for sets of eight, but sets of four not significantly differing from those for sets of eight. Also replicating Experiment 2, set variability had no effect on predictions and there was no interaction between set size and set variability. Finally, prediction task order had no effect on predictions.

When the numbers were presented sequentially, predictions underestimated the set means for all set sizes, which matched the Experiment 2 results. However, unlike Experiment 2, there was a significant overall interaction between set size and set variability. Specifically, as set size increased from four to six, predictions for high variability sets became closer to the set means as predictions for low variability sets became further from the set means. Also unlike Experiment 2, set size did have a significant effect on participant predictions, with predictions for sets of four significantly differing from those for sets of six and sets of eight. Set variability did not have a significant effect on participant predictions, which replicated Experiment 2. Finally, prediction task order had no effect on predictions.
For Experiment 3, there were no significant differences between predictions made from sets of simultaneously presented numbers and those made from sets of sequentially presented numbers. This was unlike Experiment 2, which found such a difference for sets of six numbers. There may be a few reasons why the Experiment 3 results did not fully replicate the Experiment 2 results. One reason is that the Tobii® software available to the experimenter did not allow randomization of the order of Experiment 3 prediction task trials. This lack of experimental control may have resulted in fixed response patterns across participants that are not as indicative of their actual performance. In regards to Experiment 2, the randomization of the prediction task trials may be more indicative of actual participant performance. Another reason, which is only relative to the Experiment 3 sequential presentation task analyses, is that the analyses for the Experiment 2 sequential presentation task analyses were underpowered. Therefore, it could be that the Experiment 3 results, which were based on only one more than the exact recommended amount of participants from a power analysis, are more indicative of actual performance.

**Summary for Child Predictions**

When the numbers were presented simultaneously, predictions underestimated the set means for sets of four numbers, but overestimated the set means for sets of six and eight numbers. There was a significant overall interaction between set size and set variability. Specifically, as set size increased from four to six, predictions for high variability sets became further from the set means as predictions for low variability sets became closer to the set means. Set size had a significant effect on participant predictions, with predictions for sets of four significantly differing from those for sets of
six. Set variability did not have a significant effect on predictions. Neither working memory, nor prediction task order had an effect on predictions.

When the numbers were presented sequentially, predictions underestimated the set means for sets of four and six numbers, but overestimated the set means for sets of eight numbers. There was a significant overall interaction between set size and set variability. Specifically, as set size increased from six to eight, predictions for high variability sets became significantly closer to the set means as predictions for low variability sets became significantly further from the set means. Set size had a significant effect on participant predictions, with predictions for sets of eight numbers significantly differing from those for sets of four and six numbers. Set variability did not have a significant effect on predictions. Neither working memory, nor prediction task order had an effect on predictions. There were also no significant differences between predictions made from sets of simultaneously presented numbers and those made from sets of sequentially presented numbers.

**Adult v. Child Predictions**

Whether the numbers were presented simultaneously or sequentially, there were no significant differences between adult and child predictions.

**Child Strategy Summary**

Addressing the main research question for Experiment 3, the strategy reports provided evidence that children do infer the means and variances of number sets when estimating set exemplar predictions from sets of simultaneously and sequentially presented numbers. Specifically, participants indicated that they averaged set numbers
and processed variance information (i.e., assessing set range via biggest and smallest set member, looked at the 100s column) via their open-ended reports for both sets of simultaneously and sequentially presented numbers.

Addressing the second research question for Experiment 3, the simultaneous and sequential presentation of numbers had differing influences on child prediction strategies. Simultaneous presentation prompted more rounding, particularly rounding the 10s column to the 100s column, and looking for the biggest and smallest numbers in the set. Sequential presentation prompted more averaging, looking at the 100s column of the numbers.
CHAPTER V

DISCUSSION, IMPLICATIONS, AND RECOMMENDATIONS

Strategic Inferences

The results supported that both adults and children infer set means and variances via strategic estimation of exemplar predictions from sets of eight numbers or less, regardless of whether the set numbers were presented simultaneously or sequentially. However, children exhibited a more rudimentary understanding and use of descriptive statistics and averaging strategies than adults. Children used no specific statistical terms in their strategy reports and all but one used inexact phrasing to imply they were trying to average the numbers (e.g., they described trying to find a number “in between” or “in the middle of” the set numbers) rather than saying they “tried to average” the set numbers as many adults described. Both children and adults reported various combinations of two strategies, all but one (from a child report) of which involved averaging.

It is important to point out that, as predicted, adult and child numerical prediction strategy patterns corresponded to adult and child number set comparison strategy patterns (Masnick & Morris, 2015; Morris & Masnick, 2008; Morris et al., 2014). Children reported using many more strategies than adults and their strategy use varied much more between set presentation types than the adult strategy use did (see Table 3). Important differences in usage were seen among the strategies used by both the children and adults. Trying to average the set numbers seemed to be the most effective strategy and adults
utilized it most often for both simultaneously and sequentially presented numbers. Conversely, children utilized three other strategies more often than trying to average the numbers for both simultaneously and sequentially presented numbers, half of which were strategies adults did not use at all. Also unlike the adults, children utilized strategies that were either not related to inferring the means or variances of the sets (e.g., looking at only the last few numbers in the set, adding numbers) or that did not seem to be related to the task goal of making a prediction about the set (e.g., subtracting or counting numbers). Adults did not use any strategies that were unrelated to the prediction task goal.

In line with the clear differences in adult and child strategies for number set predictions, children’s predictions were further from the set means for all set sizes of the simultaneously presented numbers (see Figure 18) and for the largest set sizes (i.e., six and eight numbers) of the sequentially presented numbers (see Figure 19). However, somewhat surprisingly, none of the differences between child and adult predictions were statistically significant. Perhaps no statistical differences were found between adult and child predictions because most children employed the same type of averaging strategies as the adults did, but they employed them in a less efficient manner (i.e., less often or in concert with strategies unrelated to the task goal).
Figure 18. Mean Experiment 2 and 3 prediction z-scores for sets of simultaneously presented numbers by set size. Note. Error bars represent SE of mean.

Figure 19. Mean Experiment 2 and 3 prediction z-scores for sets of sequentially presented numbers by set size. Note. Error bars represent SE of mean.

An even more surprising result was that having higher working memory capacity made no difference in how close adult or child predictions were to the set means.
Perhaps higher working memory capacity did not make a difference in adult predictions due to the use of averaging strategies by most participants, which appeared to be the most efficient strategies, making up for most processing capacity limitations. As for child predictions and working memory capacity, the analysis to investigate this relationship was statistically underpowered. However, particular idiosyncrasies from the children’s reported strategies indicated that working memory capacity might be a factor. For instance, one child noted on their strategy reports for processing sets of eight numbers that it “was more difficult to focus” (sequential version) and that they might not be “as accurate because there was a lot more numbers to read” (simultaneous version). These statements suggest that this child was well aware of the encoding and working memory limitations that may hinder their processing of a large amount of numbers. From the sequential version results, this child’s meta-memory awareness was correct, being that children’s prediction were furthest from the set means, by far, for sets of eight sequentially presented numbers. Therefore, further investigation of children’s numerical predictions and the influence of working memory capacity is needed.

**Shifts in Strategy Use: Simultaneous Presentation**

Even though children’s predictions were further from the set means than adult’s predictions, the descriptive patterns of child and adult predictions were identical as set size increased from four numbers to six numbers (see Figure 18). Predictions for sets of four systematically underestimated the means, whereas predictions for sets of six systematically overestimated the means. This change was accompanied by a statistically significant effect of set size, although only children’s predictions were also related to a
set size and set variance interaction. Based off the more comprehensive adult strategy evidence (i.e., response times and self-reports), a probable shift in strategy from processing four numbers to six numbers is evident. Frequency ratings from Experiment 1 show that adults looked at the 100s column increasingly more often, and looked at the 1s column increasingly less often, as set size increased from four to six numbers. Experiment 1 response times also became faster as set size increased from four to six numbers, implying that shifting more focus to the 100s column made processing more efficient. Self-reports from the Experiments 1 and 2 provided additional support that the adults focused more on the 100s column for sets of six numbers than for sets of four numbers. The increase in focus on the 100s column from sets of four to sets of six also corresponded to the change from underestimation of predictions for sets of four to the overestimation of predictions for sets of six. This implies that more focusing on more general set information (i.e., the 100s column) over the more specific set information (i.e., the 1s column) lead to overestimation of number set exemplars.

Shifts in Strategy Use: Sequential Presentation

The descriptive patterns of adult and child predictions implied dissimilar shifts in strategy as set size increased from six to eight numbers (see Figure 19). Adults systematically underestimated the means across set sizes, with predictions for sets of six numbers being significantly closer to the set means than those for sets of four, but not for sets of eight. Children underestimated the means for sets of four and six, but overestimated the means for sets of eight. This corresponded with their predictions for sets of six being significantly closer to the set means than those for sets of eight, but not
for sets of four. Therefore adults and children exhibited directly opposite effects. Specifically, adults’ predictions were furthest from the means for the smallest sets, whereas children’s predictions were furthest from the means for the largest sets.

One possible explanation for the difference in the adult and child prediction patterns is their opposing difficulties with set variance. For adult predictions, there was a significant interaction between set size and set variability between sets of four numbers and sets of six numbers. This interaction appears to be due to adults’ considerable underestimation of high variability sets of four numbers (see Figure 15), which corresponds to their predictions being furthest from the means for sets of four numbers (see Figure 19). It is possible that adults had the most difficulty with high variability sets of four numbers because they employed more equal focus on the more general set information (i.e., the 100s column) and the more specific set information (i.e., the 1s column) as the Experiment 1 frequency ratings indicated adults did for simultaneously presented sets of four numbers. This would mean that the processing strategies they employed for sets of four numbers included more of a balance between individuation and approximation than the strategies they employed for sets of six numbers. Modifications to the current studies are recommended below to allow future research to address this idea.

For child predictions, there was a significant interaction between set size and set variability between sets of six numbers and sets of eight numbers. This interaction appeared to be due to children’s relatively immense overestimation of low variability sets of eight numbers (see Figure 17), which corresponds to their predictions being furthest
from the means for sets of eight numbers (see Figure 19). A reasonable speculation for why children had the most difficulty with low variability sets of eight numbers is not readily apparent. Future research will be needed to shed light on this issue. An idea is to assign one group of children to make predictions from only sets of high variance sets and another group of children to make predictions from only sets of low variance sets and then compare the similarities and differences in each group’s strategies.

**Educational Implications**

The fact that children exhibited a more rudimentary understanding and use of averaging strategies than adults is not very surprising, as 4th grade teachers in Ohio are not required to teach their students about averaging sets of numbers (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). This means that most of the 4th grade students who participated were probably unfamiliar with formal statistical terms or the procedure for averaging sets of numbers. However, the fact that children employed inexact averaging strategies, without the aid of prior averaging instruction, supports that intuitive number approximation is not limited to counting, addition, subtraction, and magnitude comparisons (Dehaene, 2009; Feigenson et al., 2004; Gilmore et al., 2007) and may extend to estimating predictions. Therefore, number set prediction may be another mathematical ability based on representations in the ANS (Dehaene, 2001; Lerner, Dehaene, Spelke, & Cohen, 2003; Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999). In addition, along with Morris & Masnick’s (2008) number set comparison studies and Brezis et al.’s (2015) number set mean judgment
studies, the current studies provide evidence that non-symbolic number set processing ability underlies symbolic number set processing ability.

The current studies also provided evidence that children are less adept at number set variance processing than adults, suggested by the differences in adult and child variance inference strategies. The self-reports indicated that children used set variance inference strategies much more often than the adults but that their strategies may have been less effective than those of adults. For instance, children reported looking for a pattern in the numbers much more often than adults, and many pointed out that they were looking to see if the numbers were getting “bigger or smaller”, which no adult reported doing. Looking to see if the numbers were getting bigger or smaller implies that some children were trying to guess the next number in an increasing or decreasing sequence, which would produce a number far from the set average. In contrast, adults who reported looking for a pattern in the numbers usually referenced looking for specific variance information (e.g., “looking for a pattern in the 100s column”) rather than an increasing or decreasing sequence of the numbers.

Although the overall differences in adult and children predictions were not statistically significant, meaning that many children performed well on the number set prediction task, there were also many children who struggled with processing the numbers sets and producing a prediction. It is of practical concern to help students of all levels of mathematical ability improve their numerical prediction skills because such training may also have a positive impact on other areas of mathematical learning.
The current studies can provide some general guidance for prediction instruction to be applied in 4th grade classrooms.

Whether or not a child struggled with the prediction task, each child was able to articulate their strategies for processing the sets of numbers. Consequently, numerical prediction instruction should include teachers discussing strategies with students. The current studies also provided evidence that children have intuitions about using statistics to make their numerical predictions even though their teachers are not required to teach them how to do so (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). Therefore, training to improve 4th grade children’s numerical prediction estimation strategies should first include introducing them to statistical terms and the procedure for averaging numbers. A few specific instructional sequences that can follow such training are described below.

One, if we begin teaching children the more efficient estimation strategies employed by adults, their averaging intuitions should enable them to quickly learn and employ those strategies. The following instructional sequence is for teaching children to use more efficient strategies. The teacher can present children with an exemplar prediction task similar to the one used in the current studies. The teacher can have each child make their predictions and then ask some of them to share their strategies with the class. Next the teacher can explain how they went about making their prediction and set up a compare and contrast diagram with their strategy and other children’s strategies to illustrate the similarities and differences between all the strategies. Then, as a group, the class could apply each strategy to make number set exemplar predictions and see which
ones produce predictions closest to the set means. Using this approach, it would also be expected children would not only embrace the use of more efficient strategies, but also discard the use of any strategies at odds with the goal of numerical exemplar prediction.

Two, because children are less efficient in their processing of number set variance, as we teach children numerical estimation strategies, a focus should be on the most effective ways to infer set variance. The following instructional sequence is for teaching children more efficient variance processing strategies. The teacher can present children with any task that requires the averaging of numbers, and hence the inference of variance. First, the teacher can ask the students which column of numbers is most helpful in ascertaining the range of the number set. The teacher should also remind the students to pay attention to that column for each number in the set. Then, the teacher can ask the students what patterns they see in that column. If the students identify an increasing or decreasing pattern, the teacher can then discuss how thinking about that pattern actually leads one away from making a prediction based on the set average and range.

**Limitations and Modifications**

Although the current studies added to the previous literature regarding adult numerical prediction, there are still few studies on the topic. In addition, the current studies were just the beginning of research on children’s numerical prediction. Besides collecting more data to support that approximate number set averaging strategies are inherent as early as by the time children reach 4th grade (10 years of age), future studies should also begin to investigate how well younger children utilize averaging and variance processing strategies. A suggested first step to these ends is to conduct improved
versions (as described in the section below) of the current studies with another group of adults and 4\textsuperscript{th} grade students, as well as with a group of 3\textsuperscript{rd} grade students.

**Insufficient Statistical Power**

Based on the previously discussed power analysis, 35 participants were recommended for a medium effect with an estimated power of .75 at an alpha value of .05. For Experiment 2, 37 adult participants completed the study, but with the removal of outliers, the analysis for the sequential version of the number exemplar task was ultimately underpowered. This may be the reason why not all of the Experiment 2 adult sequential presentation task results were replicated by the Experiments 3 sequential presentation task results. Adult participant recruitment was more limited for Experiment 2 than Experiments 1 and 3, as the available pool of participants had more studies as options to complete to receive course credit. Yet, more data should have been collected to ensure statistical power would be achieved for all the analyses.

In addition, after the removal of outliers only 28 child participants were left to be part of the Experiment 3 working memory analysis, leaving that analysis underpowered. The lack of statistical power is also a reason to proceed with caution when discussing the results of the Experiment 3 child working memory analyses. Child participant recruitment was not limited, but with many children being unable to complete the working memory task due to various circumstances, the number of participants who could be included in the final analysis was lower than expected. Future exemplar prediction studies should, if possible, perform a check for outliers after collecting data.
from 35 participants to determine how many more participants to collect data from in order to avoid statistically underpowered analyses.

**Inadequacies Common to All Experiments**

Each of the three experiments may have suffered from the same two confounds. One, the experiment set size trial blocks were always presented in the same order, that is sets of four numbers, then sets of six numbers, then sets of eight numbers, rather than being ordered at random. This may have created a practice effect, at least in part accounting for predictions being consistently closer to the means for sets of eight numbers than for sets of four numbers. To address this, future exemplar prediction studies should include randomization of the set size blocks. If the prediction patterns from the current studies are replicated, a practice effect can be ruled out.

The second confound across all three studies was that the amount of time the sets were presented, and hence the amount of time participants had to study the numbers, increased along with set size (i.e., 4 s for four numbers, 6 s for six numbers, and 8 s for eight numbers). Therefore, participants had more time to study the larger sets. It is possible that participants did not use all the time given to study the larger sets, spending some of the intended study time actually estimating their answer. This may explain why, across the Experiment 1 and 2 simultaneously presentation prediction tasks that measured response time, the time it took participants to produce their predictions decreased as set size increased. However, a very recent set of studies provides some evidence that response times would still decrease as set size increases even if less time study time was provided for larger sets of numbers. Brezis et al. (2015) asked participants to average
sets of sequentially presented 2-digit numbers, with each set number being presented for 500ms or less and set size trials being fully randomized, and found that responses times for sets of four numbers were always slower than those for sets of eight numbers. This mirrors the results from the sequentially presented numbers of 3-digit numbers presented in Experiment 2. The same pattern held for simultaneously presented sets of 3-digit numbers in Experiments 1 and 2. It is possible that rather then due to excess study time, response times decreased as the set size of 3-digit numbers increased because processing larger sets induces participants to use more efficient, and hence faster, averaging strategies (Brezis et al.; Cravalho et al., 2013). To test this, future exemplar prediction studies should institute a 500ms per set number study time limit along with randomization of the set size trial blocks. Then, if the response time patterns from the current studies are replicated, further support will be provided that larger sets prompt an increase in the use of averaging strategies.

It is also recommended that future studies solicit confidence judgments after each prediction. Prior studies have shown that confidence judgments are related to various aspects of number set judgments such as set size, set variance, response times, and strategies (Cravalho et al., 2013; 2014; Irwin & Smith, 1956; Irwin et al., 1956; Masnick & Morris, 2008; Morris et al., 2014; Morris & Masnick, 2015). Therefore, confidence ratings may help explain aspects of participant predictions such as why response times decreased as set size increased, why participants underestimate or overestimate their predictions, or how set variance influences processing strategies.
Inadequacies Common to Experiments 2 and 3

Due to differing reasons, fewer exemplar prediction trials were included in Experiments 2 and 3 than in Experiment 1. In the case of Experiment 2, 20 trials per set size block were used, rather than 30 trials per block, as was the case for Experiment 1. Fewer trials were used for Experiment 2 as a limited amount of number sets had been created and they needed to be split between the two versions of the exemplar prediction task. For Experiment 3, only 14 trials per block were used in an effort to shorten the amount of time it would take to complete each version of the exemplar prediction task. The timing of the task became a concern of the experimenter because the teachers of the 4th grade participants voiced apprehension over how much class time a child would miss participating in the experiment. Shortening the blocks to 14 trials brought the total amount of time to participate in the experiment to a level that was more acceptable to the teachers. Having fewer trials per each exemplar prediction task could have been a reason that Experiments 2 and 3 did not replicate the entire adult Experiment 1 and 2 results. Future exemplar prediction experiments with adults should alleviate this confound by including the same amount of trials per each set size block across multiple experiments. Having fewer trials per block may have also hindered the strategy use of the child participants. It could be that 14 trials is not enough to develop a consistent strategy, hence why children reported many more strategies, many of which were not relevant to the task, than the adults. Any future exemplar prediction studies with children should make every effort to include at least 20 trials per block to give the participants enough
trials to develop a more coherent strategy for processing the amount of numbers per each set size.

The strategy report format was also altered for Experiments 2 and 3. The frequency-rating portion of the strategy report from Experiment 1 was removed from Experiment 2 mainly because one of the strategy descriptions (i.e., “Try to figure out the average.”) was believed to have possibly alerted the adult participants to the strategy of trying to figure out the set average. It was thought that this might have unintentionally altered the Experiment 1 open-ended strategy reports. However, the Experiment 2 open-ended strategy reports supported that this was not the case, as those adult participants reported trying to average the set numbers even more so than the Experiment 1 participants. Although this may have been reason to include the frequency ratings in the Experiment 3 strategy reports, there were other reasons for leaving them out. Another reason the frequency-ratings were removed from Experiment 2 and 3 was because the Experiment 1 ratings mirrored the open-ended descriptions from Experiment 1, without providing as much detail, making them somewhat redundant. Also, the frequency ratings were removed from Experiments 2 and 3 to lessen the length of the experiments so that both versions of the exemplar predictions task could be completed in the allotted time per participant. Finally, the frequency ratings were removed from Experiment 3 because this experiment was to include eye tracking as a measure of participant strategy instead.

In hindsight, there was a downside to removing the entire frequency-rating portion of the strategy reports from Experiments 2 and 3. Even though the ratings for “whole number” strategy descriptions (see Table 1, strategies 1-6) didn’t change much
across set size, the descriptions pertaining to which specific digit of each number, and
hence which columns of the numbers, the participants looked at, proved to be more
informative. Such digit/column frequency ratings could have also been triangulated with
the eye fixation data collected during Experiment 3, if that data had not been
compromised. For these reasons, future exemplar prediction studies should include
modified versions of frequency ratings portion of the strategy reports from Experiment 1.
For the simultaneous version of the exemplar prediction task, participants should be
asked to provide ratings for the following three strategy descriptions: (a) “looked at the
100s column of the numbers”, (b) “looked at the 10s column of the numbers”, and (c)
“looked at the 1s column of the numbers”. These descriptions reflect that when able to
look at all the set numbers at one time, and with them being order vertically, participants
have been found to scan the three columns of presented 3-digit numbers (Cravalho et al.,
2014; Morris et al., 2014; Morris & Masnick, 2015). For the sequential version of the
exemplar prediction task, participants should be asked to provide ratings for the
following three strategy descriptions: (a) “looked at the 1\textsuperscript{st} digit of the numbers”, (b)
“looked at the 2\textsuperscript{nd} digit of the numbers”, and (c) “looked at the 3\textsuperscript{rd} digit of the numbers”.
These descriptions more properly reflect that when able to look at only one set number at
one time, participants are confined to processing a digit rather than a column of digits.
These data could then be triangulated with eye tracking data if that is collected with
future experiments.

For Experiments 2 and 3, participants completed the simultaneous and sequential
versions of the exemplar prediction task separately. The main intention of each
experiment was to investigate performance on each task without the influence of the other. However, the separation of the two versions only allowed for an indirect comparison of the within-subject data. If future studies are motivated to directly compare predictions from sets of simultaneously and sequentially presented numbers, then it is recommended that the two versions of the task be combined. For instance, the order of each set size block of trials could be randomized and within each block half the trials could be with simultaneously presented numbers and the other half could be with sequentially presented numbers, with the trial order being randomized. This would allow the experimenter to include presentation type as a variable in one repeated measures ANOVA, rather than having to conduct separate presentation type analyses as was done for Experiments 2 and 3.

**Inadequacies of Experiment 3**

In hindsight, there were planned changes to the adult portion of Experiment 3 that had an unforeseen negative impact. One change from Experiments 1 and 2 was that adults did not complete the ABCD working memory task (cf. Was & Woltz, 2007). This was done since the focus of Experiments 1 and 2 was to investigate working memory and number set predictions, whereas the focus of Experiment 3 was to investigate attention and number set predictions. With the addition of eye tracking to Experiment 3, as the measure of attention, time was needed for calibration and data management per each version of the exemplar prediction task. Therefore, the removal of the working memory task was to facilitate the inclusion of eye tracking in an efficient manner given the timeframe to run an adult participant.
Another change from Experiments 1 and 2 was that adults did not complete self-strategy reports as part of Experiment 3. The reason for this change was again due to the focus on eye tracking and measuring attention. Being that the two groups of adults from Experiments 1 and 2 had already filled out corroborating strategy reports, it was the intention of the researchers to see if objective eye-tracking evidence for participant strategies also corroborated those subjective participant self-reports. However, in retrospect, it would have been more valuable to collect both eye-tracking and self-report strategy data from the same group of adults and then corroborate that data rather than corroborating data across groups. Initially, the child participants from Experiment 3 were not asked to report on their strategies because it was reasoned the objective eye fixation patterns would be more useful measure than their self-reports. It was thought that the children might not be able to articulate their strategies as well as the adults could. In the case of both the children and adults, it was also difficult to incorporate an open-ended strategy report at the end of each set size block (as was done with Experiments 1 and 2) given the format limitations of the exemplar prediction task programmed for use with the Tobii® T-60XL eye tracking monitor. However, after all of the Experiment 3 adult data collection and some of the child data collection had been completed without the collection of strategy self-reports, it was decided to begin asking the remaining child participants to report on their strategies after completing each version of the exemplar prediction task, rather than to report on their strategies after each block.

There are no recommendations regarding the integration of a working memory task and strategy self-reports with future eye-tracking versions of the exemplar
predictions tasks, because it is going to be recommended in the next section that eye tracking not be used for future exemplar prediction tasks. Please see that section for details.

**Sequential Version of the Experiment 3 Exemplar Prediction Task**

Due to a specific design element of the sequential version of the exemplar prediction task, some children misunderstood the goal of the task. Specifically, a fixation cross was shown before each set number as each of the set of numbers was presented (see Figure 4). Some children interpreted the fixation cross as a plus sign and as a result would add up the set numbers instead of estimating a set exemplar prediction. This design confound was not clear until the experimenter was collecting data from children during Experiment 3 because no adults reported adding numbers or interpreting the fixation cross as a plus sign when completing the sequential version of the Experiment 2 or 3 prediction tasks.

After it was discovered that some of the children were misinterpreting the fixation crosses as plus signs, extra exposition was added to the task instructions. This one sentence made it clear that the fixation crosses did not represent anything and that they were only there to make sure the child maintained their focus on the middle of the screen (where each set number would be presented). After this change in procedure, few children appeared to add the set numbers, even though a few reported addition as an initial strategy before switching to a different approach. In order to avoid this fixation cross misinterpretation in future exemplar prediction studies, only one fixation cross
should be presented before the first number of each sequentially presented set, rather than a fixation cross being presented before each set number.

A second problem was that the short presentation time of each set number slide (i.e., 1 s) caused recordings from the sequential version of the exemplar prediction task to have lower overall acuity than recordings from the simultaneous version of the task, for which the set number slides were presented for 4 s or more. Specifically, the shorter amount of time a slide was displayed, the more difficultly the Tobii® software and monitor had tracking participant eye fixations. Even though, as described earlier, other issues with the recordings preventing the coding of a substantial number of the participant eye fixations, this acuity issue would have made for an uneven balance in files to code from the two versions of the task. In addition, from watching the live feed of fixations when running participants through the sequential version of the exemplar prediction task, the experimenter did not see very much variation in fixation patterns. This means that the results of the coded recordings of eye fixations from the sequential version of the task may not have been very informative. Due to this potential lack of informative data, as well as the considerable logistics problems discussed above, it is recommended that eye tracking not be used for future versions of the exemplar prediction task.

**Summary of Suggested Modifications for Future Research**

The current studies added to the brief prior literature on adult number set predictions and extended that literature to 10-year old children. Being that even with the addition of the current studies, the number set prediction literature is still quite limited,
these suggestions will only focus on improving the next iteration of the current studies in order to pursue more evidence in support of answering the stated research questions.

Future studies of number set exemplar predictions of set smaller than eight numbers should be programmed using E-Prime® or another software that allows full randomization of the study design (i.e., set size block order, set presentation type, set variance type, trial order, and set number order). A fixation cross should be presented before each trial for sets of simultaneously presented numbers and only before the first number in the sets for sequentially presented numbers. There should be at least 20 trials per presentation type per block, and therefore at least 40 trials per set size block if one is motivated to directly compare predictions from sets of simultaneously and sequentially presented numbers. Each trial should be presented for only 500ms per set number. After each prediction trial, a confidence judgment should be solicited. After each block of trials, a strategy self-report consisting of the following two portions should be included: (a) opened strategy description prompts for each presentation type and (b) presentation type specific number place column frequency ratings. The next iteration of the exemplar prediction task, which has implemented the suggested improvements to experimental control, should be accompanied by a measure of verbal working memory capacity. It may be that the limitations of the current study were related to the lack of evidence for the influence of working memory. Finally, it is not recommended that future studies utilize eye tracking unless the researchers are motivated to triangulate such results with the strategy self-reports.
REFERENCES


*Current Psychological Research & Reviews, 3*, 5-18.


Effect of context on age and sex differences in magnitude estimation. *Journal of

Was, C. A., & Woltz, D. J. (2007). Reexamining the relationship between working
memory and comprehension: The role of available long-term memory. *Journal of
Memory and Language, 56*, 86-102.