The Roles of Inference and Associative Learning in the Construction of Mappings
Between Number Words and Numerical Magnitudes

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Psychology

by

Jessica L. Sullivan

Committee in charge:

Professor David Barner, Chair
Professor Ivano Caponigro
Professor Vic Ferreira
Professor Andrew Kehler
Professor Timothy Rickard

2014
The Dissertation of Jessica L Sullivan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_______________________________________________

_______________________________________________________________________

_______________________________________________________________________

__________________________________

Chair

University of California, San Diego

2014
DEDICATION

My thesis is dedicated to my Tanglezone Family (Adam and Alice Tinkle, Piper, Lua, Goats, Chickens, and Donkeys) and to my birth family. Without years and years of support from all of these people and non-people, this thesis would never have happened.
# TABLE OF CONTENTS

Signature Page..............................................................................................................iii
Dedication....................................................................................................................iv
Table of Contents.........................................................................................................v
List of Figures...............................................................................................................vi
List of Tables...............................................................................................................vii
Acknowledgements..................................................................................................viii
Vita..............................................................................................................................xi
Abstract of the Dissertation.........................................................................................xiii
Introduction................................................................................................................1
Chapter 1 How are Number Words Mapped to Approximate Magnitudes?..............12
Chapter 2 Inference and Association in Children’s Early Numerical Estimation.........45
Chapter 3 The Development of Structural Analogy in Number-Line Estimation........90
General Discussion.................................................................................................137
LIST OF FIGURES

Figure 1.1: Performance on the Number Matching and Discrimination tasks ..............30
Figure 1.2: Estimates by Calibration Type...............................................................34
Figure 2.1: Performance on the Number Matching and Discrimination tasks ..........64
Figure 2.2: Estimates by Calibration Type...............................................................70
Figure 2.3: Performance on the Verbal Ordering and Estimation Ordinality measures...76
Figure 3.1: Percent Absolute Error in estimation.....................................................109
Figure 3.2: Proportion of ordinal responses during estimation, plotted by age.........111
Figure 3.3: Proportion of ordinal responses plotted by estimation accuracy.............113
Figure 3.4: Proportion of ordinal responses, plotted by number being estimated......114
Figure 3.5: Estimates plotted by Distribution condition and separated by age.........116
Figure 3.6: The likelihood of a significant effect in simulated data.........................122
Figure 3.7: Five-year-olds proportion of ordinal responses, plotted by trial number....126
LIST OF TABLES

Table 2.1: Effect of calibration, separated by age and calibration type..................72
ACKNOWLEDGEMENTS

Thank you to my advisor, Dave Barner. His support, constructive criticism, and lively conversations have helped me to grow as a thinker, a person, and a psychologist. He is absolutely the best possible advisor. Thank you to the members of my committee, who have supported this work immensely and have encouraged me to expand my research line in recent years. Thank you also to Hilary Barth, without whom I never would have done psychological research.

Thank you to all of the members of LadLab, with a special thanks to Jen Audet, Karen Bejar, Amanda Chamberlain, Elizabeth Chase, Michael Sim, Ada Tong, and Shirlene Wade (for help with data collection) and to Katie Wagner (for help with MatLab code). My thesis work was supported by a Jacobs Graduate Fellowship and by an NSF GRFP, without which this work would have suffered.

Thank you to the lab coordinators who have helped facilitate this work (Jen Audet, Ellie Chestnut, and Katie Kimura) and to member of the Developmental area at UCSD. Thank you to the daycares, museums, schools, and parents who have helped provide me with participants, and, of course, to the children who have spent countless hours estimating numbers.

Thank you to my friends in the Music Department and on and around the farm (Sierra, Eowyn, Matt Patt, Leah, Matt Dooley, Greg & Jen Duckworth, Richard Border, and many others) who have helped me cultivate a fun life outside of research, and who have taught me so much about how to live life independently.
Thank you to my mom, dad, sister, aunts, uncles, and grandparents, all of whom have showered me with encouragement and love throughout my PhD, and all of whom have endured very boring conversations about my research.

And, of course, thank you to Adam Tinkle, who has sacrificed so much to help support my career and my research. As I’m writing this, he’s simultaneously cooking me dinner, watching our daughter, Alice, and arranging logistics for our impending move. Adam’s support, love, and willingness to talk to me for “just a few more minutes” while I hash out ideas about a new study or old data, has made this dissertation (and my life) much better than it would have been otherwise.

Finally, thank you to Alice Tessellate Tinkle, my amazing daughter, who has brought so much joy and data into my life.

Chapter 1, in full, is a reprint of the material as it appears in How are number words mapped to approximate magnitudes? Quarterly Journal of Experimental Psychology, 66, 389-402. DOI: 10.1080/17470218.2012.715655. Sullivan, J., & Barner, D. (2012). The dissertation author was the primary investigator and author of this paper. Permission for use of this material is granted through Taylor & Francis’ Thesis/Dissertation Reuse Policy.

Chapter 2, in full, is a reprint of the material as it appears in Inference and Association in Children’s Early Numerical Estimation. Child Development, doi: http://dx.doi.org/10.1111/cdev.12211. Sullivan J. & Barner, D. (in press). The dissertation author was the primary investigator and author of this paper. Permissions for use of this material have been obtained from John Wiley and Sons, Ltd.
Chapter 3, in full, is a reprint of the material currently under review at the *Journal of Experimental Child Psychology*. Sullivan, J., & Barner, D. The dissertation author was the primary investigator and author of this paper.
VITA

2014  Doctor of Philosophy, Psychology  
       University of California, San Diego

2010  Master of Arts, Psychology  
       University of California, San Diego

2008  Bachelor of Science, Religious Studies  
       Wesleyan University, Middletown CT

PUBLICATIONS


PUBLICATIONS UNDER REVIEW AND IN PREPARATION

Sullivan, J., & Barner, D. (under review). The development of structural analogy in

* Indicates joint first authorship
number-line estimation.


Johnson, C., Sullivan, J., Buck, C., Trexel, J., & Scarpuzzi, M. (under review). Object permanence and invisible displacement with dynamic visual occlusion in bottlenose dolphins (*Tursiops truncatus*).


ABSTRACT OF THE DISSERTATION

The Roles of Inference and Associative Learning in the Construction of Mappings Between Number Words and Numerical Magnitudes

by

Jessica L. Sullivan

Doctor of Philosophy in Psychology

University of California, San Diego, 2014

Professor David Barner, Chair

This dissertation investigates the development of mappings between the verbal number system and nonverbal modes of numerical representation.

Chapter 1 tests the proposal that adults connect number language to nonverbal number representations via two learning mechanisms: Associative Mapping and Structure Mapping. Associative Mapping in this context is an item-by-item linking of particular number words to particular nonverbal representations of number (e.g., by connecting the word twenty to nonverbal representations of ‘about 20 things’). In contrast, Structure Mapping is a holistic inference about the relation between two systems – in this case, the verbal and nonverbal number systems. Data from four tasks provide evidence that adults
rely on Associative Mappings for number words up to about 12, and Structure Mapping for larger number words. This suggests that multiple learning mechanisms aid in connecting number language to the numerical content that it represents.

Chapter 2 tests the development of children’s reliance on Associative Mapping and Structure Mapping when connecting number language to nonverbal representations of number. Do children initially rely primarily on Associative Mapping, Structure Mapping, or a combination of the two? What types of changes occur over time to children’s mappings? These data show that children rely on Associative Mappings for numbers up to about 6, and Structure Mapping for larger numbers. This suggests that (a) inferential processes like Structure Mapping are essential to even children’s earliest estimation abilities and (b) developmental changes to estimation ability in early elementary school cannot be accounted for by improvements to Associative Mappings.

Chapter 3 extends the finding that children rely on Structure Mapping to connect number language to nonverbal representations of number by asking about the types of structural inferences that children make when connecting number language to nonverbal quantity representations. This paper proposes two possible structures that could underlie SM – one based on the relative ordering of numbers and the other based on the relative distance between numbers. This paper also addresses the question of whether children make inferences about the structural relation between number language and nonverbal quantity representations on a trial-to-trial basis or over a more extended period of time.
INTRODUCTION

The Roles of Inference and Associative Learning in the Construction of Mappings Between Number Words and Numerical Magnitudes

How do we use language to describe our perceptual experiences? This dissertation uses the test case of number language as a way to understand how learners relate disparate systems of knowledge to one another. It does so by asking how the verbal number system – a precise, learned linguistic system for representing number – becomes related to the approximate and evolutionarily ancient nonverbal number system. Across three experiments, I argue that two processes guide the connection of the verbal and nonverbal number systems: item-by-item associative learning and analogically-based structural inference.

Background Information

Systems for Representing Number

Humans have access to at least two systems for representing numerical quantity: a nonverbal system and a verbal system. The nonverbal number system allows us to approximately quantify sets in the absence of counting. This system – the Approximate Number System – is present in nonhuman animals (Church & Meck, 1982; Dehaene, 1997), and in infants within their first few hours of life (Coubart et al., 2014).

The Approximate Number System (ANS) is characterized by three important properties: it is modality-independent, it is sensitive to numerical (but not other types of) quantity, and it is ratio dependent. We know that the ANS is modality-independent because numerical quantities that are presented in different domains (e.g., an array of dots
vs. a sequences of beeps) can be compared just as easily as numerical quantities that are presented in the same domain, and this is true across species and across the human lifespan (Barth, Kanwisher, & Spelke, 2003; Coubart et al., 2014; Meck & Church, 1983). We also know that the ANS actually tracks numerical quantity (and not other quantity cues like density or items size) because it successfully tracks number even in cases where other confounding quantity cues are controlled for (Xu & Spelke, 2000; Halberda & Feigenson, 2008). In fact, while confounding quantity cues like luminance, density, and item size can have small but measurable effects on numerical judgments (Hurewitz, Gelman, & Schnitzer, 2006), there is evidence that the numerical properties of sets are primary: even when explicitly instructed to make judgments about non-numerical properties of a set (e.g., by comparing the total amount of “blueness” presented after seeing two arrays of blue dots), adults are unable to ignore numerical quantity, and end up selecting the set that had a larger number of dots (Barth, 2008). Finally – and most relevant to this dissertation – we know that the ANS is ratio dependent. This means that the difficulty of discriminating two numerical quantities from one another via the ANS increases as the ratio of the number of items in the two approaches 1 (e.g., it is easier to differentiate 10 things from 30 things than 10 things from 11 things; Whalen, Gallistel, & Gelman, 1999). The acuity of the ANS increases over the developmental timespan, with typical acuity standing at a 1:3 ratio for newborns (Coubart et al., 2014), a 1:2 ratio for 6-month-olds (Xu & Spelke, 2000), a 3:4 ratio for preschoolers, and a 7:8 ratio for adults (Halberda & Feigenson, 2008). Thus, while the ANS is never able to represent exact numerical quantities, it is sufficiently powerful that a typical adult can differentiate e.g., 49 things from 56 things, even without counting.
While the ANS is useful throughout the developmental timespan, within the first few years of life, children also gain access to the verbal number system. The verbal number system allows us to talk about numerical quantity using language. Children in the preschool years can recite a count list (e.g., *one, two, three*) and use counting to determine the number of items in a set (e.g., Wynn, 1990). They can also use number language even in cases where they haven’t counted, or in approximate ways (e.g., “There were 100 people there!”).

A premise of this research is that – despite the fact that the format of the ANS is radically different from the format of number language – the verbal and nonverbal number systems become related to each other in the first few years of life. Evidence that the verbal number system becomes mapped onto the ANS comes from several sources. First, the very fact that we can use number language to label quantities without counting them suggests a connection between the ANS and the verbal number system. For example, children as young as four can use number language to label quantities in the world even without counting (Barth, Starr, & Sullivan, 2008; Le Corre & Carey, 2007; Lipton & Spelke, 2005), and evidence from ERP studies suggest that 3- to 5-year-old children are sensitive to mismatches between verbally and nonverbally presented numbers (e.g., if they hear “six” to label a set of 3 objects; Pinhas, Donohue, Woldorff, & Brannon, 2014).

Even when a task doesn’t explicitly require participants to link the verbal and nonverbal number systems together, there is evidence that activation of one system may lead to spontaneous activation of the other. For example, judgments of Arabic numerals (e.g., “which is bigger, 6 or 7?”) show ratio-dependent signatures (in that judgments are...
faster when the ratio of the quantities described by the numeral are smaller; Moyer & Landauer, 1967), even though the tasks themselves do not require accessing nonverbal number representations. It has been claimed that this pattern of performance arises because participants holistically process symbolic number language, and then rapidly and automatically convert the symbolically represented number “to an internal magnitude code” like the ANS (Dehaene, Dupoux, & Mehler, 1990; Dehaene, Dehaene-Lambertz, & Cohen, 1998).

Similarly, both verbal (e.g., “27”) and nonverbal (e.g., 27 dots) representations of number activate identical regions within the intraparietal sulcus, or IPS (Fias et al., 2003), and rTMS stimulation to the left IPS disrupts the processing of both verbal and nonverbal numerical quantity, and does so in a way that is ratio-dependent (Cappalletti et al., 2007). While evidence of overlapping regions of neural activity for verbal and nonverbal number judgments is certainly not sufficient – on its own – to demonstrate that the ANS and verbal number system are related to one another, when paired with the behavioral results described above, we have strong evidence that the ANS becomes related to the verbal number system. The relevant research question, therefore, is not whether the verbal number system becomes mapped onto the nonverbal number system, but rather how such mapping process might work.

Mappings Between the Verbal onto Nonverbal Number

In this dissertation, I describe and test two possible learning mechanisms that could guide the mapping of verbal number language onto nonverbal representations of number: Associative Mapping and Structure Mapping. Both of these mechanisms have long histories in philosophy and psychology, and both have been shown to be useful
learning mechanisms in other domains. As I have argued elsewhere (Sullivan, 2011; Sullivan & Barner, 2012; 2014), understanding the relative roles of these two mechanisms will not only improve our understanding of language learning more generally, but will also help us to understand the emergence of early numerical (and perhaps mathematical; Siegler & Ramani, 2009; Siegler & Booth, 2004) competence.

**Associative Mapping**

If we connect the verbal number system to our nonverbal number system via Associative Mapping (AM), then each number word gets associated – on an item-by-item basis and via experience – with a particular state of the ANS. Thus, via experience with particular word-magnitude pairings (e.g., hearing “seven students” in the presence of a class of 7 students; hearing “seven dollars” in the presence of 7 dollar bills), we might slowly form connections between approximate nonverbal representations of a given numerical quantity (e.g., ‘about 7 things’) and a number word (e.g., “seven”). By this view, number word mappings are independent from one another such that the mapping for the word *seven* is formed via experiences with 7 things, and the mapping for the word *fourteen* is formed via experiences with 14 things.

While most research on the development of number knowledge avoids the question of how number words become mapped to the ANS (claiming only that there is such a mapping), when any mapping mechanism is specified, it is typically an Associative Mapping mechanism. For example, the AM view of number word mappings is made explicit in the model of symbolic number acquisition proposed by Verguts & Fias (2004) as described by Ansari (2008): “symbolic [numerical] representations are
learned through the simultaneous presentation of symbolic and non-symbolic inputs”. ¹

Neuroscientists have continued to adopt the AM model, attempting to test the neural correlates of number language mapping via studies of adults who were “trained to associate novel symbols with nonsymbolic numerical magnitudes (arrays of dots)”, an endeavor that – by default – assumes that number word mappings are typically formed via AM (Lyons & Ansari, 2009). Developmental researchers have also converged on the Associative Mapping view when describing how the verbal and nonverbal number systems become linked to one another. For example, Lipton & Spelke (2005) propose two alternatives for how the verbal and nonverbal number systems might become mapped to one another. The first is that “children might learn the mapping by forming direct associations between individual number words and nonsymbolic numerosity representations...they might learn that the word ‘hundred’ refers to approximately the number of marbles that would fill a vase”. The second alternative is that counting skills might precede the formation of mappings, such that “As they produce each number word [when counting], they may associate with the set of objects counted thus far”. In both cases, the process described is what I – and the authors – would call “associative mapping”. And, as I have argued elsewhere (Sullivan, 2011), across several subfields

¹ Ansari (2008) rightly notes that while current data are consistent with this associative learning account, there are alternative explanations of these data. In the end, Ansari (2008) neither endorses nor dismisses the AM account. It is also worth noting that Verguts & Fias (2004) also propose that any association between verbal and nonverbal number may be mediated by a third process of place coding.
within numerical cognition, AM appears to be the default assumption for how number words become related to the ANS.  

As will be discussed in detail in Chapters 1 and 2, there are several shortcomings to the Associative Mapping view, which suggest that Associative Mapping may not be the only learning mechanism that we recruit when connecting our verbal and nonverbal number system. An alternative mechanism – Structure Mapping – is described below and tested in Chapters 1, 2, and 3 of this dissertation.

*Structure Mapping*

Structure Mapping is an inferential process by which a single link is formed between two systems on the basis of their shared structure (Gentner, Anggoro, & Klibanoff, 2011; Gentner & Namy, 2006; Carey, 2009; Gentner, 1983). Outside of the domain of number, a child might use structure mapping to label a large vase “dad”, a medium-sized vase “mom”, and a small vase “baby” by noticing a shared structure (increasing size) across families and vase sizes (Ratterman & Gentner, 1998). Within the domain of number, a child might notice structural similarities between the verbal and nonverbal number systems, and make a holistic inference about the connection between these two systems on the basis of their similar structures (Cantlon, Cordes, Libertus, & Brannon, 2009; Carey, 2009; Izard & Dehaene, 2008; Thompson & Opfer, 2010). For instance, a child might notice that both the verbal and nonverbal number systems are ordered: the count list is an ordered list such that items that come later refer to larger quantities, and it is possible to place numerical quantities in order from smallest to

---

2 See also Barth, Beckman, & Spelke (2008) and Huttenlocher, Jordan, & Cohen Levine (1994) for discussions of accurate and automatic translation between the ANS and the verbal number system that implicitly assumes something like AM.
largest. Or, the child might notice that both number systems can encode relative distance: twenty comes twice as far in the count list as ten does, and a set of 20 objects contains twice as many objects as does a set of 10. In this way, each mapping is mutually constraining: What a child knows about one number word’s relation to the nonverbal number system will influence what they know about all other number words. While the experiment presented in Chapter 3 will discuss which structural relations (e.g., relative ordering; relative distance) children use when forming a SM, Chapters 1 and 2 will focus on understanding the developmental trajectory of the use of SM in connecting the verbal to the nonverbal number system.

**Current Directions**

This dissertation contains 3 chapters, each of which focuses on better understanding the learning mechanisms used to connect language to nonlinguistic input. In Chapter 1, I test the relative roles of Associative Mapping (AM) and Structure Mapping (SM) in the formation of adults’ mappings between verbal and nonverbal number representations. In Chapter 2, I extend the findings in Chapter 1 by testing the developmental trajectory of the use of AM and SM in supporting children’s mappings between verbal and nonverbal number representations. In Chapter 3, I test – in greater detail – children’s use of SM, and outline a view of numerical estimation that highlights (a) the role of analogy in constructing estimates; (b) the role of working memory in constraining children’s use of SM; and (c) the role of trial-to-trial calibration in shaping SM. In the discussion, I describe the theoretical, empirical, and educational impacts of this work, and briefly discuss future directions.
References


Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition, 86*, 201-221. DOI: 10.1016/S0010-0277(02)00178-6,


Huntley-Fenner, G. (2001). Children’s understanding of number is similar to adults’ and rats’: numerical estimation by 5-7 year olds. *Cognition, 78*, B27-B40. DOI: 10.1016/S0010-0277(00)00122-0


Evidence for the Involvement of the Approximate Number System in Preschoolers’ Processing of Spoken Number Words. *Journal of Cognitive Neuroscience.* doi: 10.1162/jocn_a_00631.


Chapter 1

How are Number Words Mapped to Approximate Magnitudes?
Abstract

How do we map number words to the magnitudes they represent? While much is known about the developmental trajectory of number word learning, the acquisition of the counting routine, and the academic correlates of estimation ability, previous studies have yet to describe the mechanisms that link number words to nonverbal representations of number. We investigated two mechanisms: associative mapping and structure mapping. Four dot array estimation tasks found that adults’ ability to match a number word to one of two discriminably different sets declined as a function of set size and that participants’ estimates of relative large, but not small, set sizes were influenced by misleading feedback during an estimation task. We propose that subjects employ structure mapping for linking relatively large number words to set sizes, but rely chiefly on item-by-item associative mappings for smaller sets. These results indicate that both inference and association play important roles in mapping number words to approximate magnitudes.
How does language represent human numerical knowledge? Do we identify the referents of words like *twelve* and *fifty-seven* by associating them, item-by-item, with sets in the world? Or are the referents of numerals determined primarily by inference and logical relations between words? In recent years, experimental psychologists have uncovered non-linguistic foundations of number in human infants, cross-cultural abilities to reason about number in absence of counting ability, and numerical capacities in non-human animals (for review, see Feigenson, Dehaene, & Spelke, 2004; Dehaene, 1997). These discoveries have laid the groundwork for studying how words become related to numerical representations of sets, and thus for addressing how inferential and associative mechanisms contribute to the development of number words in language.

By some accounts, number words get their content in part via a mapping to non-linguistic representations of number provided by the Approximate Number System, or ANS. The ANS, which is found in human adults, infants, and a host of other animals, allows individuals to represent and compare numerical magnitudes nonverbally and is governed by Weber’s Law (for review, see Dehaene, 1997). Weber’s Law states that the discriminability of two perceptual stimuli is determined by the ratio of their magnitudes. Because Weber’s law governs the ANS, judgments of numerical magnitudes are also ratio-dependent, and exhibit the signature of scalar variability—error increases in proportion to the size of the sets being represented. As a result, it is equally easy to differentiate 6 vs. 7 objects as it is to discriminate 600 vs. 700. Verbal estimates of sets also demonstrate the signatures of scalar variability, providing strong evidence that mappings exist between the verbal number system and ANS representations (e.g., Izard
While it is clear that we can map number words onto ANS representations of numerical magnitude, surprisingly little is known about the nature of these mappings, and descriptions of possible mechanisms are rare in the literature. It is therefore not well understood what roles associative and inferential processes play in linking number words to ANS representations.

In the present paper, we explored the roles of two candidate mechanisms for linking number words to nonverbal numerical representations: Associative Mapping (AM) and Structure Mapping (SM). According to the AM hypothesis, as humans accumulate experience with number words, they form item-specific associations between individual words (e.g., ten) and corresponding ANS representations of numerical magnitudes (e.g., approximately 10 objects). This view posits that each number word is mapped individually to a particular state of the ANS through experience with particular word-magnitude pairings. Because each word is linked on an item-by-item basis to ANS representations, mappings should also be relatively independent of one another.

In contrast, according to the SM hypothesis, the verbal and nonverbal number systems become mapped onto each other on the basis of the structural similarities between them (Carey, 2009; Gentner, 1983; Gentner & Namy, 2006). Structure mappings support reasoning across many contexts – like language learning and understanding spatial relations – by relating the structure of one domain to a corresponding structure in a second domain (for discussion, see Gentner & Namy, 2006). In the case of number, the ANS provides an ordered set of magnitude representations, while the count list provides a corresponding ordered set of numerals. Thus, to support estimation, individuals may recruit SM to link their two number systems on the basis of their shared ordinal structure.
Contrary to AM, SM predicts that individual number word mappings should be non-independent: the content of one number word should depend on the content of all other known numbers.

Predictions of the AM and SM hypotheses

To understand the predictions that AM and SM make, it is instructive to consider how they might be deployed online, while doing estimates. By the AM hypothesis, each number word is associated with a particular ANS representation, such that mappings are independent of one another. This predicts that in estimation experiments, an adjustment to the mapping for the word twenty (perhaps via feedback) won’t automatically affect other mappings (e.g., the mapping for the word forty). Also, because mappings constructed via AM are linked directly to particular ANS representations, adults should use unique number words for each discriminably different magnitude (e.g., a different number word should be assigned to sets of 10 and 20). Thus, to the extent that two magnitudes are reliably discriminable, subjects should not assign them the same label (e.g., adults should almost never guess that a set of 40 contains twenty, since they discriminate sets in a 2:1 ratio with very high accuracy). Thus, two predictions of the AM hypothesis are that AMs should not be easily influenced by feedback regarding other number words, and that, when shown two sets, adults should have little difficulty choosing which set corresponds to a provided number word, so long as they two sets are readily discriminated by the ANS (e.g., they should easily judge whether fifty applies to a set of 50 or a set of 100).

In contrast, according to the SM hypothesis, the verbal and nonverbal number systems become mapped onto each other on the basis of their structural similarity (Carey,
2009; Gentner, 2010; Gentner & Namy, 2006). Because SM involves the association of two global structures, the mappings for individual numbers should be non-independent, and instead should be defined in relation to one another. This hypothesis makes several clear predictions about estimation performance. First, a subject’s estimate for one set should constrain their estimates for other sets (such that bigger sets receive bigger estimates). Also, when a subject’s response for a given quantity is changed via feedback (i.e., calibrated), responses for other quantities should also change correspondingly, potentially beyond the range of predicted by error in the ANS (see Tversky & Kahneman, 1974, for a conceptually related discussion of anchoring and adjustment in numerical judgments). Finally, because SMs are constructed based on context dependent inference, it is possible that, across different situations, subjects will fail to provide distinct labels for discriminably different sets (e.g., a subject might label a set of 40 as forty in one context, and call a set of 20 forty in another context).

Evidence for AM and SM in estimation

Although past research has not described estimation in terms of these two mechanisms, it seems likely that both AM and SM must play some role in linking numerical magnitudes to nonverbal representations of number, and there is some evidence consistent with this view. For example, it is well known that adults are almost errorless when estimating sets with up to 4 items, and that these estimates remain accurate across different experimental contexts (Mandler & Shebo, 1982). Such evidence is consistent with the view that mappings up to four are supported by AM. Further evidence for AM comes from the child estimation literature. After learning how counting represents number, many children are able to make rapid estimates for small sets, but not
for slightly larger sets within their counting range (Le Corre & Carey, 2007). This suggests that these children do not initially possess abstract structural knowledge of the relationship between verbal and nonverbal representations of number, and therefore that their mappings for smaller numbers are item-specific. Taken together, these studies of adults and children suggest that AM may play a role for at least some small number words. However, these studies have failed to test whether (1) these mappings remain independent of other mappings later in life, and (2) whether associative mappings might be acquired for numbers beyond three or four.

There is also evidence consistent with the view that SM supports estimation for at least some numbers. In one study, Izard and Dehaene (2008) provided explicit miscalibration to adults by mislabeling a visually presented set (e.g., by calling 30 dots “twenty-five”). After calibration, participants shifted their estimates not only for the calibrated quantity, but also for many of the other quantities tested. Consistent with SM, these findings demonstrate that changes to the mapping for one word can affect the mappings of other words. The SM hypothesis also predicts that estimates made early in an experiment should constrain later estimates, and recent work has shown that subjects’ estimates are influenced by the magnitude of the very first number they are asked to estimate (Sullivan, Juhasz, Slattery, & Barth, 2011). For example, participants who made estimates for relatively small numbers at the start of an estimation task systematically shifted their estimates for all subsequent trials, relative to subjects who began with larger numbers. Consistent with SM, these data suggest that subjects dynamically adjust their mappings between number words and magnitudes in the course of doing experiments, and that estimates for one number word affect those for all others. Finally, there is recent
evidence that elementary schoolers can be trained to analogically extend information about familiar numbers to an unfamiliar number range (Thompson & Opfer, 2010). They can also incorporate accurate estimation feedback regarding one number to improve estimation performance for other numbers (Opfer & Siegler, 2007). Consistent with SM, these developmental studies show that children can use information regarding one range of numbers to modify their estimates for another range of numbers (but see Barth, Kanjlia, Slusser, Garcia, & Chase, 2011 for an alternative interpretation of these data).

Although these previous studies are consistent with the view that both AM and SM can be used to guide estimation, they do not adequately differentiate their respective roles. We see four broad possibilities left open by past work for how these two systems might interact. One possibility is that AM guides mappings for most number words, and that the role of inference in constructing estimates is restricted to fine-tuning existing mappings. By this view, number words are associatively linked to particular ANS states, and effects of calibration therefore only emerge when participants use SM to make small, local adjustments to their mappings. A second possibility is that, among adults, SM guides all estimates, even for small numbers that are initially acquired using associative mappings. This possibility is predicted by Izard and Dehaene’s (2008) affine-transformation model, which posits that all mappings undergo a “global transformation” when miscalibrated so that “all numerosities are calibrated at once” (p. 1244). Third, it is possible that AM is limited to very small numbers in the subitizing range, like 1-4, and that estimates for all larger numbers are supported by SM. This view is consistent with evidence that children acquire one to four associatively, and also with reports that SM is used to support estimates for large numbers (Izard & Dehaene, 2008). Finally, a fourth
possibility is that AM and SM interact to support estimates for both large and small numbers, but that associative mappings are strongest for small and frequent numbers, and play a declining role as numbers grow larger and are used less frequently to refer to perceptual sets.

*Limitations of previous work*

Although past studies have shown effects consistent with SM, they have not set out to test how AM and SM combine to support estimation. These previous studies have shown effects of SM only for relatively large numbers (Izard & Dehaene, 2008; Thompson & Opfer, 2010), and only with small manipulations (Izard & Dehaene, 2008; Sullivan et al., 2011), such that its role relative to AM is impossible to determine. For example, in their study, Izard and Dehaene (2008) analyzed subjects’ performance for large numbers but not for smaller ones, leaving open whether SM was used for small numbers, as predicted by their model (but not predicted by models where AM plays any role in estimation). Also, they provided relatively modest miscalibration, resulting in small mean differences between calibration conditions. When participants were induced to overestimate, they mapped the number word *thirty* to sets averaging 27.4, while those calibrated to make accurate estimates mapped *thirty* to 31.5. Critically, 27.4 and 31.5 differ by almost a 9:10 ratio—a discrimination threshold that has not been reliably documented in adults. While participants who were induced to underestimate showed somewhat larger effects of calibration, the reported differences were still within the range of error previously documented in studies of the ANS, making it difficult to determine which, if any, mappings were anchored by AM.
Taken together, previous studies leave open the relative roles of AM and SM in supporting numerical estimation. Although developmental studies point to a possible role for AM for small numbers, and adults studies show preliminary evidence that SM is deployed for large numbers, almost nothing is known about how these mechanisms interact to generate estimates – i.e., whether AM extends beyond 4 (interacting with SM to generate larger estimates), and whether SM extends to small numbers. Although it may seem likely that each mechanism must play some role in estimation, this assumption is currently only that—a best guess about what likely links language to perceptual representations of number.

The present study

Understanding the mechanisms that support estimation is critical to multiple fields of inquiry, including the acquisition of number word mappings (Barth, Starr, & Sullivan 2009; Lipton & Spelke, 2005; Le Corre & Carey, 2007), how mappings are represented in the brain (Ansari, 2008; Dehaene & Changeaux, 1993; Piazza, Pinel, Le Bihan, & Dehaene, 2007), and how best to train mappings in children (a significant predictor of math success in early childhood; Siegler & Booth, 2004; Siegler & Ramani, 2009). Currently, the literatures that discuss these topics do not consider the nature of the mechanisms that support estimation, or what the relative roles of such mechanisms might be.

The present study was the first to explicitly test the roles of AM and SM in linking the ANS to the count list. To contrast these possibilities, we tested the extent to which SM and AM guide array estimation for relatively small vs. relatively large number words in a series of four tasks.
EXPERIMENT

To test the relative roles of AM and SM in verbal estimation, we conducted an experiment with four within-subjects measures that probed for evidence of associative and structure mappings between number words and ANS representations across a wide range of numerosities.

The first pair of tasks that participants received involved making verbal and nonverbal numerical judgments about two sets of dots. In this pair, the critical test was the Number Matching task, in which subjects were asked to match a provided number word to one of two sets that differed by a 1:2 or 3:4 numerical ratio. Unlike in previous studies using similar methodologies, we did not provide a familiarization phase (Mundy & Gilmore, 2009), so that we could assess participants’ mappings in the absence of feedback. Performance on this task was compared to participants’ performance on a Number Discrimination task. The Number Discrimination task used identical stimuli to the Number Matching task, but participants only had to decide which of two sets was larger (a purely nonverbal, ANS-based judgment). Comparing performance on these tasks allowed us to assess whether participants possess veridical and stable mappings between number words and magnitudes. We predicted that if participants have strong AMs, then they should use these mappings to restrict judgments in the Number Matching task. For example, to the extent that a subject can reliably discriminate sets of 20 and 40, they should never assign the number word twenty to a set of 40 dots, provided that AM guides mappings for twenty. However, if AM does not support estimates for twenty, then subjects may not reliably apply it to sets of 20 rather than sets of 40.
In the second set of tasks, participants saw single arrays one-at-a-time and had to estimate how many dots they saw. In the Calibrated Estimation task, participants made dot-array estimates after receiving highly misleading information about the range of set sizes to be presented (e.g., that the largest set would be 75, 375, or 750, when it was in fact 350). These data were compared to data from an earlier Uncalibrated Estimation task, in which the same participants received no information about the sets they would see. We predicted that feedback should not influence participants’ estimates for quantities that have strong AMs, but should influence estimates for all magnitudes with SMs. This is because number words mapped using AM should be linked on an item-by-item basis to magnitudes, and therefore should have relatively independent mappings, while SMs are constructed in relationship to each other.

Methods

Participants

Thirty adults from the UCSD community participated for course credit. One additional participant was excluded from analyses for failure to complete all tasks. All participants provided written informed consent.
Procedure

Participants were seated approximately 40 cm from a 27” Mac OSX computer screen and completed 4 computerized tasks. Half of the participants completed the Number Matching task first, and half completed the Discrimination task first. All participants then completed the Uncalibrated and Calibrated Estimation tasks. For all tasks, presentation time for the stimuli was brief, preventing the use of counting.

Number Matching. In this task, participants heard a number word, saw two dot arrays flash sequentially on a computer screen, and judged which array best matched the word. Stimuli were sets of red dots presented on a black screen for 400 ms, backwards masked with random noise for 100 ms. On each trial, sets differed in numerical magnitude by either a 1:2 ratio or 3:4 ratio. Sets were matched for item size on half of the trials and for total occupied area on the other half (Dehaene, Izard, & Piazza, 2005), and comparisons ranged from small (4 vs. 8) to large (370 vs. 740). Participants saw 35 comparisons (see Figure 1), each a total of four times. On half of the trials, they were asked about the larger of the two sets; on half of the trials, the first set presented was correct.

Numerical Discrimination. This task served as a within-subjects control for the Number Matching task to ensure that participants could discriminate the quantities presented. The stimuli and procedure were identical to the Number Matching task, except that participants indicated which set contained more dots (instead of matching a word with a set). Again, on half of the trials, the first set was correct.

Uncalibrated Estimation. Participants saw sets of dots and estimated their numerosities, recording their estimates using the numeric keypad on a computer.
keyboard. Stimuli were sets of red dots on a black screen. Fifteen numerosities were presented 36 times each [8, 12, 20, 35, 60, 80, 95, 120, 150, 180, 200, 240, 275, 300; 350]. These magnitudes were selected so that all estimates would require use of the ANS (e.g., Feigenson et al., 2004). Each numerosity was matched for both item size (15 trials) and total occupied area (15 trials) with each other numerosity presented (Dehaene, Izard, & Piazza, 2005), and non-numerical properties of the sets were otherwise varied for the remaining 6 trials. There were 540 stimulus arrays in total, which were randomly presented to subjects across the Uncalibrated and Calibrated Estimation tasks, for a total of 270 trials in each.

**Calibrated Estimation.** Stimuli and instructions were identical to those in the Uncalibrated Estimation task, except participants were given information about the largest set they would see. In order to distinguish between the SM and AM hypotheses, we made two critical methodological changes from previous studies that used similar paradigms (e.g., Izard & Dehaene, 2008; Shuman, unpublished thesis). Whereas participants in past studies were shown an array and told that it contained $x$ dots (where $x$ was either an accurate or inaccurate number word label), in the present study we did not mislabel visible arrays. Instead, we told subjects, “the largest set you will see is $x$”. In this way, we ensured that any influence of feedback was not because participants constructed new associative mappings, but was due purely to an inference about the Structure Mapping relation. A second difference was that we provided much more extreme calibration than in previous studies, in order to test the strength of Associative Mappings throughout the number line. Although the largest set that participants saw was 350 in all conditions, we manipulated the degree of miscalibration between subjects.
Participants were told at the start of the task that the largest set they would see was either 75 (N=10), 375 (N=10), or 750 (N=10).

Results

Number Matching and Discrimination

If participants have strong AMs between individual number words and approximate magnitudes, then they should be able to use these mappings to guide the labeling of sets in the Number Matching task. For example, if the number word twenty is associatively mapped to a mental representation of ‘about 20 things’, then participants should never match the word twenty to an array that is discriminably different from ‘about 20’ (e.g., 40). Because of this, performance should be relatively high on the Number Matching task for magnitudes that are associatively mapped. Said differently, it should be nearly as easy to map the number word twenty onto a set of 20 (and not 40) as it is to nonverbally discriminate sets of 20 and 40. However, for words that lack strong AMs, subjects may exhibit relatively low accuracy, compared to their performance for the same numbers on the purely perceptual number discrimination task.

To test this, we first analyzed the difference in accuracy on the Number Matching task compared to the Discrimination (baseline) task. All 30 participants performed worse on the Number Matching task than on the Discrimination task, and a paired-samples t-test comparing mean accuracy for each stimulus pairing on the Number Matching task to mean accuracy for the identical pairing on the Discrimination task revealed that this trend
reached significance (all $p < .05$) for 21/30 participants (binomial probability $p < .01^3$).

This difference in accuracy between the Number Matching and Discrimination tasks is surprising if participants used AM to guide their mappings—in order to perform worse on the Number Matching task than on the Discrimination task, participants had to attach the wrong number word label to one of two discriminably different sets.

Notice that this cannot be explained by a tendency to underestimate the magnitude of sets, or by the possibility that participants simply had non-veridical mappings for many number words. For example, one might propose that participants have strong AMs, but that they systematically underestimate the quantities presented, resulting in poor Number Matching accuracy. On this account, if participants associatively mapped the word *forty* to the magnitude 80 (underestimation), they should perform better when, after seeing a set of 40 vs. 80 items, they are asked to match *eighty* than when asked to match *forty*. This is because when asked to map *eighty* to 40 vs. 80, even though the participant (erroneously) thinks that the larger of the two sets would be better described by *forty*, the only sensible choice would be to map *eighty* onto the bigger set (and not onto an even smaller set!). However, when asked to match *forty* to either 40 or 80, the participant could rely on his existing and inaccurate mapping, thus mapping *forty* onto 80 and getting the trial incorrect.

To assess this possibility, we analyzed accuracy on trials where the correct answer was the larger of the two sets, compared to when it was the smaller of the sets. We found

---

3 For brevity, we will use cutoffs of $p < .05$ and $p < .01$ for this and all binomial analyses. Because we would expect a handful of participants to show significant effects by chance alone due to the large number of individual analyses, binomial analyses help gauge whether particular patterns of performance are more prevalent in a population than would be expected by chance.
no evidence that participants systematically underestimate (or overestimate). Only 5/30 participants performed better when the foil set was larger than the requested number (predicted by underestimation), whereas 8/30 participants performed better when the foil set was smaller than the requested magnitude (overestimation). The dominant pattern was to show no effect of foil size (17/30). Critically, all three of these subgroups showed the same basic pattern of results overall, suggesting that the effects reported above (and below) were not driven by any particular subgroup. Taken together, these results indicate that the decline in accuracy on the Number Matching task, relative to the Discrimination task, cannot be explained solely by inaccurate (but potentially associative) mappings.

The next analysis assessed the possibility that increased task demands led to lower performance on the Number Matching task relative to the Discrimination task. It also tested a critical question necessary to distinguishing the roles of AM and SM. If Number Matching is simply harder than Numerical Discrimination, then there should be a task-related main effect, such that performance on the Number Matching task is always lower than performance on the Discrimination task. However, if Number Matching performance varies because only some numbers are supported by AM, then the task effect should interact with set size, and performance should be high for mappings supported by AM, and lower for mappings supported by SM. To foreshadow results for the Calibration task, described below, we found that miscalibration had no effect for numbers 12 and smaller, a result that is consistent with the use of AM for those quantities. If this analysis of the Calibration finding is correct, then performance on the Number Matching task should be best for these same quantities, with diminishing
performance for larger numbers. To address this, we compared accuracy on the Number Matching task to Discrimination accuracy for each stimulus pairing.

Participants performed significantly worse on Number Matching than Discrimination for 20/35 comparisons presented. Critically, participants did not perform significantly worse for any of the eight comparisons containing magnitudes smaller than 15 (Dunnett’s mean comparison, all $p > .05$). Accuracy also did not differ between Number Matching and Discrimination for four of the five largest comparisons presented. However, we do not take this as evidence for AM for two reasons. First, there was no evidence that participants performed well on other large number words (see below for more analyses related to this), and there was no a priori reason to think that the largest number words would be more likely to be mapped via AM than other slightly smaller (but more frequent) words like fifty and one hundred. Second, this pattern of performance is highly consistent with an SM account: Participants may have made inferences about which number words and sets seemed to be the largest ones presented, and relied on these inferences in order to adopt strategies like “when I hear an unusually large number word, I will select the larger of the two sets” or “when I see an unusually large set, I know it should be paired with an unusually large number word” to increase performance on these trials.\(^4\)

\(^4\) In addition studies not reported here, we varied the range of number words being judged, and replicated the effect that performance on the largest few comparisons tested is relatively high (regardless of the magnitude of these larger comparisons). Also, identical methods used to test children find the same pattern (Sullivan & Barner, in prep). Both results support the idea that improved performance on the largest trials is due to an inference about the largest numbers tested in a given task, and not to AMs.
Figure 1.1: Performance on the Number Matching and Discrimination tasks at (a) 1:2 ratio and (b) 3:4 ratio. Data points are means; error bars are SEM. X-axis is the smaller of the two sets being compared, since knowledge of the mappings for the smaller of the two sets is sufficient to guarantee accuracy for that comparison.
In order to further probe the effect of set size on accuracy on the Number Matching task, we used the LME4 package in R to construct a Linear Mixed Model (LMM) of accuracy on the Number Matching task by set size, with ratio as a fixed factor, and subject as a random factor (Bates & Sarkar, 2007; R Development Core Team, 2010). For this and all subsequent LMMs, we report the coefficient and standard error, as well as the $p$-value estimated from Markov chain Monte Carlo simulations. This analysis revealed an effect of Ratio (1:2 vs. 3:4; $\beta = -.012, SE = .002, p < .001$) and an effect of Set Size ($\beta = -.0004, SE = .00009, p < .001$), but no interaction ($p > .75$). Consistent with previous research, participants had greater difficulty with comparisons at a 3:4 ratio than a 1:2 ratio. However, because the ratio of the number of objects in each comparison remained fixed across the range of magnitudes tested, the increase in error due to set size cannot be attributed to perceptual properties of the sets tested. Instead, participants had greater difficulty matching number words to larger sets relative to smaller ones, consistent with the possibility that small number words have stronger associative links to approximate magnitudes. These results suggest that participants have relatively strong AMs for small number words, but not for larger ones.

**Estimation and Calibrated Estimation**

Before conducting analyses, we excluded all responses of 0 and 1, as well as all responses more than 10 times larger or smaller than the presented numerosity, as these were likely to be typing errors on the part of participants (N = 106/15,540). Additionally, we removed outliers by excluding all data points more than 3 SD from the mean of each participant’s estimate of each presented set size (N = 313/15,540).
Overall, participants in the current study provided estimates that were linearly related to the presented set size, and were influenced by misleading feedback (miscalibration). In this case, being miscalibrated means being induced to either over- or under-estimate, relative to the Uncalibrated Estimation task, on the basis of feedback given at the beginning of our Calibrated Estimation task. A LMM\(^5\) of participants’ estimates predicted by Set Size and Calibration type (Calibrated, Uncalibrated) with Subject as a random effect revealed significant effects of Set Size (β = .47, SE = .008, \(p < .001\)) and Calibration (β = -6.15, SE = 2.26, \(p < .01\)), and a significant interaction of Calibration and Set Size (β = .116, SE = .012, \(p < .001\)). Participants were influenced by misleading feedback, but only for relatively large numbers.

Notice that this asymmetry cannot be explained by the fact that representations of larger numerosities exhibit more error (as predicted by scalar variability). Specifically, if estimates of larger numbers are -- regardless of Calibration condition-- less accurate and more variable than estimates of smaller numbers, then calibration could appear to have a bigger effect on larger numbers than smaller numbers. To rule this out, we computed a normalized measure of the effect of calibration—one that accounted for each participant’s estimation accuracy in the absence of calibration. We found identical results to those reported above. For each set size tested, we calculated a normalized estimation measure for each participant:

\[
\frac{\text{Mean Estimate}_{\text{Calibrated}} - \text{Mean Estimate}_{\text{Uncalibrated}}}{\text{Mean Estimate}_{\text{Uncalibrated}}}
\]

\(^5\) Note that, while this model is a linear model, consistent with past research on estimation (e.g., Izard & Dehaene, 2008; Stevens, 1957), our data are well fit by Steven’s power model. However, as shown in previous work on miscalibration, testing the effects of calibration on a linear model of estimation is appropriate (Izard & Dehaene, 2008).
This resulted in a positive score for participants who were induced to overestimate and a negative score for those who were induced to underestimate; the more this number deviated from zero, the stronger the effects of calibration. This measure accounts for the error attributable to the ANS itself (since this should be constant across the different calibration conditions), allowing us to assess the effect of calibration as a proportion of change in estimation behavior. Using Dunnett’s Mean Comparison, we used this normalized measure to test effects of Calibration type on estimation behavior for each numerosity tested. There were no differences between Calibration types for the numerosities 8 or 12 (ps > .15), but there were effects for all larger magnitudes (all ps < .05). Even with a normalized measure of the effect of Calibration, participants’ performance only differed as a function of Calibration for magnitudes 20 and larger.

To explore the influence of feedback at an individual level, we next regressed each participant’s estimates onto Set Size, by Calibration type. Twenty-four out of 30 participants showed a significant effect of calibration: 9/10 participants who were calibrated to 75, 8/10 who was calibrated to 375, and 7/10 who were calibrated to 750 (significantly more participants showed an effect of calibration than could be accounted for by chance; binomial p < .01 for each calibration type). Of the 24 participants who were influenced by calibration, 21 demonstrated a significant interaction between Calibration and Set Size (binomial p < .01), indicating that calibration influenced estimation differently as a function of Set Size.

---

Additional analyses on the non-normalized measure of estimation performance revealed identical effects: calibration influenced estimates for all magnitudes larger than 12.
Figure 1.2: Estimate by calibration type. Data points are means.
Specifically, participants were less influenced by misleading feedback for smaller sets, and were more influenced for larger sets. Note that there was variability across participants in the point at which Calibration began to have an effect. Nearly half ($n=11$) of the 24 participants who were influenced by calibration showed effects of calibration for magnitudes of 20 or smaller, yet two participants remained uninfluenced by calibration for magnitudes up to 150. This suggests that there is nothing special about the number 12 as a cut-off point between AM and SM: instead, the strength of AM declines as a function of numerical magnitude, and may differ across participants.

**Discussion**

This study provides the first empirical test of how two distinct mechanisms (associative mapping and structure mapping) combine to link number words and approximate magnitudes. For the smallest numbers we tested (up to ~12), participants reliably matched number words to arrays, and were resilient to misleading calibration. This suggests that, at least for relatively small and familiar set sizes, adults form strong associative mappings between number words and ANS representations. However, for larger sets, adults struggled to correctly match number words to arrays and were highly influenced by misleading calibration when estimating, making it unlikely that associative mappings play an important role in relating larger number words to magnitudes. Calibration affected performance across a wide range of set sizes, and most participants were influenced by calibration for sets smaller than 100. Similarly, on the Number Matching task, accuracy declined rapidly as a function of set size and was low for larger numbers, despite the fact that the discriminability of sets in each comparison remaining constant across trials. These results provide strong evidence that associative mappings are
either extremely weak or entirely absent for larger numbers, and that structure mappings play a primary role in guiding estimation for most numbers, beginning with numbers as small as 15 or 20.

The fact that subjects were accurate at Number Matching and unaffected by calibration for numbers up to 15 is consistent with the idea that strong associative mappings (AMs) are present not only for one, two, and three, but also for larger numbers, with mappings growing gradually weaker as a function of set size. Thus, whereas previous research has argued that calibration should cause a global transformation of all mappings (Izard & Dehaene, 2008), we found that many of the smallest set sizes that we tested were un-influenced by misleading feedback. This result is consistent with previous reports that, as the size of numbers increases, our experience hearing and interacting with them decreases correspondingly (Dehaene & Mehler, 1992). Because associative learning requires mapping a number word to a magnitude via experience with particular word-magnitude pairings, this reduced experience with larger number words may result in weaker AMs.  

Of course, while the resilience of small numbers to misleading feedback suggests that SM has little influence on these mappings, it doesn’t guarantee that all such mappings are purely associative. It is also possible, for example, that associative mappings are strong only for the very smallest numbers (e.g., 1-4) and that participants visually chunk larger arrays into sets of 4, making estimates for slightly larger sets by

---

7 Note that, as previously mentioned, the number 12 as a cutoff between AM and SM is not likely to be special. Although 12 was the largest number to be unaffected by calibration group-wise, individual subjects did show variability in performance. Also, the effect of miscalibration did not decrease abruptly after 12, but simply fell below the threshold of significance.
using some form of multiplicative strategy (for discussion of perceptual chunking, see Atkinson, Francis, & Campbell, 1976; Mandler & Shebo, 1982; van Oeffelen & Vos, 1982). Past studies have shown that, when visual arrays of dots are grouped into small sets, subjects are highly accurate when estimating relatively small quantities, but less aided in their estimation of larger sets (e.g., Frank & Barner, 2012). This fact alone suggests that visual chunking might extend the range of AM. However, the same studies (and many other in the estimation literature) fail to find evidence of a chunking advantage for small sets when they are not visually arranged into smaller groups, suggesting that chunking may only be used when made possible by the structure of visual arrays.

Likewise, our study found that subjects rarely made accurate estimates even for the smallest sets we tested. Thus, although it remains possible that subjects make subtle, undetectable, use of chunking when visual arrays are randomly configured, there is currently no evidence to support this hypothesis.

Critically, whether or not participants used chunking when estimating, our main conclusions remain the same: the role of AM in estimation is relatively minor and is restricted to the smallest numbers we tested, while SM operates for most larger magnitudes. According to the SM hypothesis, most number words lack strong associative links to the ANS (e.g., there is no fixed and veridical link between fifty and representations of ‘about 50 things’ in the ANS). Instead, individuals construct a structure mapping that relates the verbal and nonverbal number systems on the basis of their structural similarity. As in other cases of structural (or “analogical”) mapping (e.g., in the domains of spatial reasoning and category learning), individuals may use a small set of associative mappings to license a broader structural mapping (e.g., Gentner, 2010;
Gentner & Namy, 2006). Specifically, subjects may recruit associative mappings for a handful of small numbers to construct a structure mapping for larger numbers.

While we have argued that AM operates for only small sets, it is important to note that this claim regarding the nature of language-to-magnitude mappings is separate from the question of how sets are represented non-linguistically. According to some accounts, there are two distinct systems capable of representing quantity non-linguistically – i.e., the object-file system (or “parallel individuation”), which can represent up to 4 objects in parallel, and the ANS, which can represent the approximate magnitude of any set (Feigenson et al., 2004). The distinction between these systems was not the focus of our study, since regardless of how sets are represented non-verbally, either AM or SM could still apply to sets larger than 3 or 4. Also, previous accounts explicitly predict that SM should apply to even the smallest numbers, right down to one (Izard & Dehaene, 2008). Because we were specifically interested in the mapping mechanism and not the format of non-verbal numerical representation, we restricted our manipulations to numbers well beyond the range of the object-file system, and found clear signatures of the ANS for all sets tested (i.e., scalar variability). Thus, our study assessed the mechanisms that support mappings between language and perception, and the effects we found speak to the nature of these mappings – e.g., how strong mappings were for individual words, and how easily they could be recalibrated on the basis of misleading feedback – independent of how sets are represented non-verbally.

Note that the presence of this signature is another reason to doubt that subjects estimate sets like 12 or 15 by chunking estimates supported by object files. Such a strategy would not predict the typically observed pattern of scalar variability.
Understanding the mechanisms that relate language to numerical perception is critical not only to explaining adult estimation, but also to understanding numerical development in childhood. While much has been learned in recent years about number word learning and early estimation abilities, little is known about the mechanisms by which number words get mapped to the ANS in development. For example, we know that even after learning how counting represents number, many 4-year-olds fail to map larger number words to larger sets in estimation tasks, whereas 5-year-olds do so successfully (Le Corre & Carey, 2007; Lipton & Spelke, 2005; Barth, Starr, & Sullivan, 2009). Clearly, 5-year-olds have learned something about the count system that the 4-year-olds have not—but what? The present study suggests two possibilities. First, it is possible that around the age of 5, children’s estimation abilities improve as a result of learning and refining a set of associative mappings, and that, early on, even competent estimators lack SMs. On this view, a structure mapping may only emerge after children have acquired relatively adult-like AMs (e.g., with strong associations up to 12-15). A second possibility is that children use SM to guide estimation from the beginning, and that children acquire SMs on the basis of minimal evidence (e.g., associative mappings for only the smallest number words, like 1-3). Consistent with this idea, children’s estimation abilities are predicted by their familiarity with the count list, such that higher counters are better estimators (Davidson, Eng, & Barner, 2012; Lipton & Spelke, 2005; Carey, 2009; although see Barth et al., 2009 for evidence that even poor counters may

---

9 For this to be true, these early AMs for small numbers could not involve the ANS (since previous studies suggest that the ANS becomes mapped to number words only after one, two, and three are known). Instead, early AMs might involve other systems with numerical content, such as “parallel individuation” (see Le Corre & Carey, 2007).
draw on structural knowledge of the count list when estimating). On this view, some AMs may be acquired after an SM is already in place, and may continue to change throughout development. Future studies should explore the roles of inference and associative learning to understand the sources of developmental differences in estimation ability, and how such differences relate to mathematical outcomes in the classroom, which are known to be correlated with estimation ability (Siegler & Ramani, 2009; Booth & Siegler, 2008).

To summarize, we have demonstrated how associative learning and inferential processes combine to guide numerical estimation in adults. In doing so, we have taken a first step in understanding the mechanisms that guide the mappings between number language and nonverbal numerical representations. Future work is required to refine and model the particular ways in which these two learning mechanisms interact throughout development and into adulthood. Characterizing these mechanisms will, in turn, help us to understand how language and nonlinguistic perceptual systems interface, and how inferential capacities combine with experience to form conceptual knowledge of number.
Chapter 1, in full, is a reprint of the material as it appears in How are number words mapped to approximate magnitudes? *Quarterly Journal of Experimental Psychology, 66*, 389-402. DOI: 10.1080/17470218.2012.715655. Sullivan, J., & Barner, D. (2012). The dissertation author was the primary investigator and author of this paper. Permission for use of this material is granted through Taylor & Francis’ Thesis/Dissertation Reuse Policy.
References


Frank, M., & Barner, D. (2012). Representing exact number visually using mental
abacus. *Journal of Experimental Psychology: General, 141*, 134-149. DOI: 10.1037/a0024427


Chapter 2

Inference and Association in Children’s Early Numerical Estimation
Abstract

How do children map number words to the numerical magnitudes they represent? Recent work in adults has shown that two distinct mechanisms – Structure Mapping and Associative Mapping – connect number words to nonlinguistic numerical representations (Sullivan & Barner, 2012). Here, we investigated the development of number word mappings, and the roles of inference and association in children’s estimation. We tested 58 5-to-7-year-olds, we found that at both ages 5 and 7, children possess strong item-based associative mappings for numbers up to around six, but rely primarily on Structure Mapping – an inferential process – for larger quantities. We conclude that children rely primarily on an inferential mechanism to construct and deploy mappings between number words and large approximate magnitudes.
How do number words like *seven* and *thirty-five* get linked to things in the world? When children make these mappings, do they form item-by-item associations between numerals and quantities? Or do they use inferential processes, based on their knowledge of how numbers are related to one another? Although philosophical discussions of mathematical knowledge provide strong reasons to doubt that perception alone could supply the logical meanings encoded by number words (e.g., Frege, 1884/1953; Kant, 1781), experimental psychologists have established that once children acquire such meanings, number words do eventually get mapped to perceptual representations of quantity (e.g., Carey, 2009; Gelman & Gallistel, 1978; Le Corre & Carey, 2007; Siegler & Opfer, 2003; Wynn, 1990, 1992). Surprisingly little is known, however, about the mechanisms by which linguistic and non-linguistic representations of number become linked during development, and thus what roles inference and association play in this process. Here, we explored this question, and asked how young children begin to map number words onto perceptual representations of quantity.

Humans and non-humans alike have access to an Approximate Number System (ANS) for representing numerical content (for review, see Dehaene, 1997). Consistent with Weber’s law, this system supports the nonverbal comparison of sets on the basis of their numerical ratio. For example, sets that stand in a 1:2 ratio are equally easy to discriminate whether the sets are relatively small (e.g., 5 vs. 10) or relatively large (e.g., 500 vs. 1000; Barth, Kanwisher, & Spelke, 2003; Brannon & Terrace, 2000; Feigenson, Dehaene & Spelke, 2004; Whalen, Gallistel, & Gelman, 1999). The ANS is used to represent numerical content in early infancy, and by at least 6-months of age, infants can reliably discriminate quantities at a 1:2 ratio (e.g., Xu & Spelke, 2000). The acuity of this
system grows slowly over development well into the teenage years, and typically converges on a ratio of about 7:8 or higher in adults (Halberda & Feigenson, 2008; Halberda, Mazzocco, & Feigenson, 2008).

The ANS eventually becomes linked to the verbal number word system, probably sometime after children begin to acquire the logical meanings of these words (Le Corre & Carey, 2007). Evidence for this comes primarily from studies of estimation. In typical dot-array estimation experiments, participants see a series of rapidly flashed dot arrays, and then label each array with a number word (Atkinson, Francis, & Campbell, 1976; Barth, Starr, & Sullivan, 2009; Frank & Barner, 2012; Huntley-Fenner, 2001; Izard & Dehaene, 2008; Le Corre & Carey, 2007; Lipton & Spelke, 2005; Mundy & Gilmore, 2009). As the number of items in an array grows, the degree of error in the participant’s estimate grows, too (e.g., Whalen, et al., 1999). Critically, this pattern of error, like that found in studies of nonverbal numerical comparison, can be described by Weber’s law, suggesting that numerical estimation requires linking number language to the ANS (for related evidence, see Dehaene, 1989; Moyer & Landauer, 1967; see also, Vul, Barner & Sullivan, 2013). Evidence for such a link is found not only in adults, but also in children as young as 4 years of age (Holloway & Ansari, 2009; Huntley-Fenner, 2001; Le Corre & Carey, 2007). By this age, but probably not before, children have constructed some sort of rudimentary mapping between number words and ANS representations. Once children have begun to form these mappings, their estimates remain relatively inaccurate for several years, but improve slowly over development (Barth et al., 2009; Berteletti et al., 2010; Booth & Siegler, 2008; Ebersbach et al., 2008; Lipton & Spelke, 2005; Mundy & Gilmore, 2009; Siegler & Opfer, 2003). Little is known, however, about the learning
mechanisms which underlie the formation and refinement of mappings between the ANS and the verbal number system, and thus what causes these changes in estimation ability.

In studies of adults, two mechanisms have been proposed to explain how the verbal and nonverbal number systems might become linked: Associative Mapping, an associative mechanism (see Lipton & Spelke, 2005), and Structure Mapping, a mechanism that relies on structurally mediated inference (Carey, 2009; Gentner, 1983; Gentner & Namy, 2006; Sullivan & Barner, 2012). Associative Mapping (AM) involves the creation of item-by-item associations between particular number words and the magnitudes they represent. For example, for a word like twenty, the creation of AMs involves associating the word twenty with a nonverbal (ANS) representation of approximately 20, via experience in the world (e.g., ‘20 students’, ‘20 crackers’, ‘20 minutes’, etc...) and by not associating it to sets of discernibly different magnitudes (e.g., “twenty” does not apply to ‘10 students’, ‘40 crackers’, or ‘60 minutes’).

Structure Mapping (SM), in contrast, involves creating a single link between the count list and numerical representations in the ANS. This link is formed on the basis of the similarity of the structures of these two systems, rather than on associations between particular numbers and sets of things in the world. For example, according to the SM hypothesis, knowledge that the number word forty comes after the word twenty in the count list guarantees that forty will always be mapped to larger quantities than twenty. This is because the count list, like the ANS, is an ordinal system of representation, and is structured such that the ordering of its symbols is predictive of the ordering of approximate magnitude representations in the ANS. Further, not only are representations in each system ordered, but the distance between symbols in the count list is predictive of
the distance between the magnitudes that they encode. Thus, a mature SM might reflect not only the relative ordering of numbers in both systems, but also their relative distance (e.g., that *forty* is twice as far into the count list as *twenty* and therefore should be mapped onto a set that is twice as large as the one labeled by *twenty*). Given these properties, according to SM, the mappings for number words are defined in relation to one another, and do not require a veridical link between, for example, *forty* and ‘40 things’ (for a related discussion of anchoring and adjustment, see Tversky & Kahneman, 1974). Crucially, the content of each number word in a system created through SM is dependent on the content of all other number words.

In order to understand the predictions that AM and SM make, consider how these mechanisms might be recruited during estimation. By the AM hypothesis, each number word is associated with a particular state of the ANS, such that mappings are independent of one another. This predicts that adjustments to one mapping (e.g., via feedback) should not automatically influence other mappings: misleading feedback about number word mappings should cause local – but not global – shifts in estimation performance. Also, on average, estimates of discriminably different magnitudes should be reliably different, because these magnitudes will almost always be mapped to different number words (e.g., estimates for sets of 50 and 100 should typically differ).

In contrast, according to the SM hypothesis, number words become related to magnitudes via a global mapping between two structures (the verbal and nonverbal number systems), and estimates are based on inferences about the ordering and distance between number words and magnitudes. Thus, number word mappings are non-independent and mutually constraining, such that an estimate for one set should constrain
estimates for all other numbers. When a participant’s estimate for a given quantity is changed via feedback (i.e., calibrated), other estimates should also change correspondingly. Also, depending on contextual factors and feedback (Izard & Dehaene, 2008; Sullivan, Juhasz, Slattery, & Barth, 2011; Sullivan & Barner, 2012), discriminably different magnitudes may not always be mapped onto different number words (e.g., a participant might provide the same estimate for a set of 50 in one context, as they do for 100 in another context).

There is evidence that adults use both AM and SM to support estimation. In one recent study (Sullivan & Barner, 2012), adults were asked to match a number word to one of two discriminably different dot arrays (e.g., map the word fifty to either 50 dots or 100 dots). Participants succeeded when the magnitude of the arrays was relatively small (10 vs. 20), but performed substantially worse for larger comparisons (e.g., 50 vs. 100), despite the fact that these comparisons differed by the same numerical ratio. Importantly, these same participants showed no effect of set size when asked to perform a numerical discrimination task, in which they judged which of the two dot arrays was larger. This pattern of findings resulted in an interaction between magnitude and task: subjects easily discriminated sets regardless of their magnitude on the discrimination task, but their success at matching number words to one of these two sets was mediated by magnitude (i.e., they were much better at mappings words to sets for smaller comparisons relative to larger ones). This result suggests that participants relied heavily on AM for small number words, but much less for larger numbers.

In a second set of tasks, adults made estimates of dot arrays after receiving misleading information about the largest quantity they would see in the experiment (e.g.,
being told that the largest set they would see is 750 when it was actually 375). Relative to a baseline estimation task, estimates made after this miscalibration were shifted for all but the smallest numbers, resulting in an interaction between magnitude and task (calibrated vs. uncalibrated estimation). Together, these two sets of tasks provided strong evidence that adults (1) relied primarily on AM for smaller number word mappings, (2) had weaker AMs for larger magnitudes, and (3) relied more heavily on SM for larger magnitudes. These findings are consistent with other studies of estimation, which have also found that information about one mapping can affect estimates for other number words (Izard & Dehaene, 2008; Shuman, unpublished thesis; Sullivan, et al., 2011).

Although these adult studies establish that both inference and association are deployed by mature estimators, they leave open how such mappings are constructed, and how AM and SM interact in early development. As a result, it is currently unknown how the mapping process begins, and whether children use inference to guide estimation from the beginning, or are initially restricted to item-based associative learning.

One previous study found evidence that children can recruit SM when estimating, but only tested older children who were already competent estimators (Thompson & Opfer, 2010). In their study, Thompson and Opfer (2010) trained elementary school aged children to recognize the structural relation between a familiar number range (e.g., 0-100) and an unfamiliar number range (e.g., 0 to 10,000). This training improved their estimation performance for the unfamiliar number range, suggesting that children used their structural knowledge of the smaller number range to guide their interpretation of the larger numbers. This suggests that when children understand how a particular part of the count list is structured (e.g., the ordering and distance between numerals), they can use
SM to guide their estimates. However, the study did not test the mechanism children use to make spontaneous, untrained, estimates, leaving open the possibility that even elementary-school aged children do not typically recruit SM to guide estimates. Also, it did not test how SM emerges earlier in development, and how it relates to AM – i.e., whether children initially rely more on AM or use SM from early in development.

In the present study, we investigated how children construct early mappings between number language and the ANS, by testing whether they begin by relying primarily on item-based associations, or instead use structure based inferences. We also investigated whether children’s reliance on these two mechanisms changes during development, as their estimation abilities improve (Siegler & Booth, 2004). We see several possible ways that these systems could interact to support estimation. On the one hand, it is possible, as argued by Lipton and Spelke (2005), that children initially rely heavily on associative mappings (e.g., perhaps by forming associative mappings for magnitudes up to 50 or 100). By this view, early mappings might be formed exclusively via associative item-specific learning, but slowly give way to SM as children learn the global relation between counting and numerical magnitude. Alternatively, it is possible that children initially rely primarily on SM to guide estimation, and acquire associative mappings slowly, as they gain experience with estimation (see Izard & Dehaene, 2008, for a model that posits only SM). A final possibility is that children acquire a small set of strong AMs – e.g., up to 10 or 12, like adults – before beginning to use SM for estimation.

While many previous studies have investigated the mechanisms by which children acquire the logical meanings of number words (beginning with one, two, three, and then
learning the principles that govern counting; see Carey, 2009, for review), no previous study has distinguished between the mechanisms that might guide children’s mappings between number words and perception. Understanding the mechanisms that guide the formation of these mappings is important for several reasons. First, estimation performance is known to be linked to a host of educational outcomes (Booth & Siegler, 2008; Siegler & Ramani, 2009; Siegler & Booth, 2004), so understanding the learning mechanisms that allow children to become successful estimators is vital to understanding the factors that drive math success. Second, estimation performance is known to improve dramatically over development (e.g., Barth & Paladino, 2011; Booth & Siegler, 2008; Siegler & Opfer, 2003; Slusser, Santiago, & Barth, 2012), yet it remains unknown whether such improvements are guided primarily by improvements to nonverbal number knowledge, verbal number knowledge, or to the mappings between them. Finally, and perhaps most importantly, testing the roles of AM and SM in supporting estimation provides a test case for investigating the fundamental question of how inference and association combine to link language to the content of the world.

EXPERIMENT

We used four within-subjects computerized measures to assess children’s knowledge of number word mappings. These measures explicitly tested the roles of AM and SM in guiding verbal estimation and were adapted from previous adult studies of these learning mechanisms (Sullivan & Barner, 2012). When considered together, these tasks address the nature of children’s early mappings by testing the roles of AM and SM in estimation. We also used two non-computerized assessments of verbal number knowledge.
First, we assessed whether mappings for particular number words were formed via strong AMs. To do this, participants completed a pair of tasks requiring judgments about sequentially presented arrays of dots. In the Discrimination task, participants decided which of two sets contained more dots. Performance on this task, which only required nonverbal number knowledge, served as a baseline for our critical measure, the Number Matching task. In this task, participants decided which of two presented sets matched a particular number word (a judgment that requires mapping the verbal number system to ANS representations of number). By comparing performance on the Number Matching task to the baseline performance on the Discrimination task, we assessed whether participants reliably map particular number words (e.g., fifty) to one of two discriminably different sets (e.g., 50 vs. 100). This allowed us to test the degree to which number words had strong associative mappings. For magnitudes that are mapped via strong AMs, performance should be relatively high on the Number Matching task, since visual arrays should activate their appropriate verbal labels directly. However, to the degree that children lack strong AMs, they should be less accurate at these judgments, even for highly discriminable quantities. Critically, if children – like adults – rely more on AM for small numbers than for larger numbers, we should find that performance on the Number Matching task remains relatively high for small number words (relative to the Discrimination task), and declines as a function of numerical magnitude for larger number words, resulting in a significant interaction between task and magnitude (for discussion, and similar data from adults, see Sullivan & Barner, 2012).

In a second set of tasks, we more explicitly assessed the role of SM in guiding early estimation performance by testing whether misleading feedback influenced
estimation performance. Participants completed two estimation tasks. In the first task, they were given no feedback about the range of numbers being estimated (Uncalibrated estimation); in the second task, an experimenter provided a single instance of calibration by telling participants the magnitude of the largest set that they would see during the experiment (Calibrated estimation). We asked whether children’s estimates were influenced by misleading feedback and therefore whether they recruited SM to support their estimates. We did this by comparing performance on the Uncalibrated Estimation task to performance on the Calibrated Estimation task. If participants rely primarily on AM for a given quantity, then calibration should have no effect, since associative mappings are independent from one another. However, calibration should have a significant effect on estimation performance if mappings rely strongly on SM, since this would indicate the use of a single data point to realign the entire structure mapping. Again, if AMs are relatively strong for small numbers but decline in strength and give way to SMs for larger numbers, we should see no effect of calibration for the smallest numbers estimated, and a larger effect of calibration for larger numbers.

In a third set of tasks, we assessed children’s verbal number knowledge, to determine how the use of different estimation strategies in the first tasks was related to children’s knowledge of the count list. This was of interest since, according to past studies, counting ability predicts certain aspects of estimation performance (Barth et al., 2009; Davidson, Eng, & Barner, 2012; Lipton & Spelke, 2005). There were two tasks. First, participants were given two counting assessments – a free count exercise (“count to 100”) and a scaffolded counting exercise (“finish this counting sequence”). The second task tested children’s understanding of the relative ordering of number words.
Participants decided which of two boxes contained more stickers after hearing, for example, that one box contained *twenty* stickers and one box contained *forty* stickers. Children who possess a strong understanding of the order and structure of the count list should succeed at this task, while those who do not will likely fail (e.g., Davidson et al., 2012; Le Corre, in prep).

**Method**

**Participants**

Thirty-two 5-year-olds (range: 5;0-5;11) and 26 7-year-olds (range: 7;0-7;11) participated. Participants were recruited from a database of interested families maintained by the psychology department at UCSD, and were compensated for travel expenses and given a small prize. Participants lived in the greater San Diego area, and were primarily Caucasian and upper-middle class. Of the 58 participants tested, 5 were excluded: for failure to complete at least 10 trials on any task in the experiment (*n* =2), due to inattention (*n* = 2), or due to computer error (*n* = 1). Data from the 7 participants who successfully completed at least one critical pair of tasks (e.g., Number Matching and Discrimination; Estimation and Calibrated Estimation; Counting and Verbal Ordering) but failed to complete all tasks were included in the relevant groupwise analyses. The remaining 46 participants (21 boys) contributed a full dataset (23 5-year-olds ranging in age from 5;0 to 5;11 and 23 7-year-olds ranging in age from 7;1 to 7;11) and were included in all analyses.

**Procedure**

Participants were tested in a quiet lab setting, after written parental consent and the child’s verbal assent were secured. Participants were seated approximately 40 cm
from a 27” Mac OSX computer screen while they completed four computerized tasks and two non-computerized tasks. The total testing time was approximately 1 hour, and participants were offered breaks between each game. Participants completed the Discrimination task first, the Number Matching task second, the Uncalibrated Estimation task third and the Calibrated Estimation task fourth. For all computer tasks, the presentation time for stimuli was brief (400 ms), preventing the use of counting. The majority of participants completed both non-computerized tasks (Counting and Verbal Ordering) after completing all four computer tasks—however, some participants completed one or both of these tasks as part of a break between computer tasks. Because task order was fixed (to prevent e.g., performance on the Calibrated estimation task from influencing performance on the other computerized tasks), it was not possible to test whether task order influenced performance. However, performance on the Discrimination and Number Matching tasks could not have been influenced by the non-computerized tasks (because they always came before these tasks), and there was no qualitative evidence of an influence of non-computerized task order on estimation (e.g., children who did the counting task before estimating were not more likely to provide sequential estimates on the estimation task than children who did the counting task after).

Numerical Discrimination. Participants saw two sets of dots, and had to decide which set had more dots. The arrays were presented on a black background and were flashed sequentially for 400 ms, and each set was backwards masked with random noise for 100 ms. On each trial, sets differed in numerical magnitude by a 1:2 ratio. Trials were presented in a fixed random order. There were 20 possible comparisons (see Figure 1), and each participant saw each comparison up to four times. Sets were matched for item
size on half of the trials and for total area on the other half (MatLab code: Dehaene, Izard, & Piazza, 2005), and comparisons ranged from small (3 vs. 6) to large (300 vs. 600). The task ended after the participant saw all 80 trials, or after 10 minutes. The color of the dots was constant within a given trial, but varied across trials (colors included red, blue, yellow, white, cyan, green, and magenta). On half of the trials, the first set contained more dots. Participants chose the set they thought was larger either by saying “first set” or “second set” or, for older children, by entering ‘1’ or ‘2’ onto the numerical keypad. This task functioned as a within-subjects control for the Number Matching task to ensure that participants could discriminate the quantities presented.

**Number Matching.** The stimuli and procedures were identical to those used in the Numerical Discrimination task, except that participants were instructed to match a given number word to one of two visually presented arrays. A numeral appeared in black font on a gray background before the onset of the trial, and the experimenter read the numeral out loud if the child did not spontaneously do so. Only after the child heard the number word did the experimenter show the arrays. On half of the trials, the numeral matched the larger of the two sets; also, on half of the trials it matched the first set presented.

**Uncalibrated Estimation.** On both this and the calibrated estimation task, participants saw a single dot array on each trial and were asked to estimate how many dots they saw. All participants provided their responses verbally, and an experimenter entered them using a numeric keypad. Stimuli were sets of dots on a black screen and were taken from the pool of arrays presented in the Discrimination and Number Matching tasks. Twenty-six numerosities were presented up to 3 times each: 3, 4, 5, 6, 8, 10, 12, 16,
Arrays were presented in one of two fixed random orders. The task ended after the participant saw all 78 trials, or after 10 minutes.

**Calibrated Estimation.** Stimuli and instructions were identical to those in the Uncalibrated Estimation task, except that participants were first told the size of the largest set they would see (Sullivan & Barner, 2012). Although the largest set that participants saw was 600 in all conditions, they were randomly assigned to be told that the largest set they would see was either 25, 75, or 750 (with approximately one-third of the children in each age group assigned to each calibration condition). While this degree of miscalibration here may seem extreme, our previous work suggests that it goes unnoticed by naïve subjects. In our adult study of calibrated estimation (Sullivan & Barner, 2012), participants who were asked whether they thought that calibration was misleading rarely reported that it was (despite sometimes being calibrated to expect a maximum of 75, when the actual maximum was 350). In the present study, only one participant (a 7-year-old) thought that the calibration was “silly”.

**Counting Assessment.** Participants were asked to guess how high they could count. For the free-count assessment, they then counted as high as they could. Each child was encouraged to continue counting until they reached 100 (e.g., Barth, et al., 2009; Davidson, et al., 2012), or until they made 8 errors, all of which were recorded by the experimenter. Participants were then given an additional scaffolded counting assessment (e.g., Lipton & Spelke, 2005). For this task, the child was asked to finish counting a sequence that an experimenter started – e.g., “7, 8, 9 – what comes next?” All participants successfully said “ten”, and then completed the remainder of the counting
assessments, which included: 16, 17, 18, ___, ___; 48, 49, ___, ___; 97, 98, ___, ___; 247, 248, ___, ___; 296, 297, ___, ___; 447, 448, ___, ___; 498, 499, ___, ___; 997, 998, 999, ___.

**Verbal Ordering Assessment.** Children’s knowledge of the ordering of the number words was assessed in this task. Participants were shown two toy gift boxes and were told that there were stickers inside each box. The experimenter then said, “This box [pointing to box on left] has X stickers in it, and this box [pointing to box on right] has Y stickers in it. Which box has more stickers?”, where X and Y were replaced with two number words that differed by a 1:2 or 3:4 ratio. The child was then asked to point to the box that contained more stickers. Participants completed 16 trials involving numbers ranging from 4 to 700—on half of the trials, the box on the left had more stickers. Trials were presented in one of two fixed random orders.

**Results**

**Number Matching and Discrimination**

Here, we asked three main questions about the role of AM in children’s estimation. First, is there evidence that children have AMs for at least some number words? Second, how high do strong AMs extend? Finally, how does children’s use of AM change over time (e.g., what changes between 5 and 7 years of age, as children become competent estimators)?

To test the role of AM in supporting children’s estimation, we asked whether performance on the Number Matching task declined significantly as a function of the numerical magnitude being tested, relative to performance on the Discrimination task.
While it was possible that the Number Matching task would be generally more difficult than the Discrimination task due to the increased demands it places on subjects, the key test of our hypothesis was whether performance on the Number Matching task interacted with magnitude, as found in our previous study of adults (Sullivan & Barner, 2012). In the current study, the ratio between magnitudes on both tasks was always held constant, at a 1:2 ratio. Therefore, any interaction due to magnitude can only be explained by the relative strength of AMs for numbers of different sizes.

For these analyses, we constructed a binomial logit model using the lmer package in R (Bates & Sarkar, 2007; R Development Core Team, 2010), predicting task accuracy from task (Number Matching or Discrimination), the smaller magnitude in any given comparison (e.g., ‘3’ for the comparison 3 vs. 6), and their interaction. For this and all other Linear Mixed Models (LMMs), subject was considered a random factor. In 5-year-olds, we found an effect of task ($\beta = 1.1$, SE = .09, $p < .0001$), an effect of magnitude ($\beta = -.001$, SE = .0007, $p < .01$), and no interaction ($\beta = -.0008$, SE = .0009, $p > .05$). Overall, participants performed worse on the Number Matching task than on the Discrimination task, and performance on both tasks declined as numerical magnitude increased. In 7-year-olds, we also found an effect of task ($\beta = 1.94$, SE = .12, $p < .0001$), and an effect of magnitude ($\beta = -.004$, SE = .0007, $p < .0001$). Critically, in these older children, an interaction of task and magnitude also emerged ($\beta = -.0023$, SE = .001, $p < .05$). Thus, like the 5-year-olds, 7-year-olds performed worse on the Number Matching task than the Discrimination task, and worse on larger magnitudes than smaller magnitudes. Also, like in adult populations (Sullivan & Barner, 2012), the effect of magnitude was mediated by task: performance declined steeply as a function of
numerical magnitude for the Number Matching task, while performance on the Discrimination task did not (see Figure 2.1).

From these data, we can conclude that, at least by the age of 7, children, like adults, rely strongly on AM for relatively small numbers and rely much less on AM for larger numbers. The results for the 5-year-olds are somewhat more difficult to interpret. Without a significant interaction between Magnitude and Task, it’s difficult to determine whether 5-year-olds recruit AM when estimating. The first possible explanation of this finding is that 5-year-olds – like adults and 7-year-olds – displayed a magnitude-based decline in accuracy on the Number Matching task (relative to the Discrimination task), but that this trend was not detected by our statistical test. This would predict that – despite a lack of interaction – a direct comparison of performance on the Number Matching task to the Discrimination task for each comparison presented (e.g., 3 vs. 6, 4 vs. 8, … 300 vs. 600) should reveal no difference in task performance for smaller comparisons, and large differences in performance for larger comparisons. A second possibility is that children relied strongly on AM for both small and large numbers, such that analyses of individual comparisons should reveal no difference in performance on the Discrimination and the Number matching task for all comparisons tested. A third possibility is that children lacked AMs for any of the magnitudes we tested. If this were the case, large differences between performance on the Discrimination and Number Matching tasks should emerge for all comparisons tested.
Figure 2.1: Accuracy on the Number Matching task (solid line) and Discrimination task (dashed line) for (A) 5-year-olds and (B) 7-year-olds. All comparisons were presented at a 1:2 ratio. The smaller number in each comparison is represented on the x-axis (e.g., 3 denotes the comparison of 3 vs. 6). Error bars denote the SEM.
To assess these possibilities, we used Dunnett’s Mean Comparison – an analysis that corrects for multiple comparisons – to test whether accuracy differed on the Number Matching task relative to the Discrimination task for each comparison tested. For 5-year-olds, accuracy on these two tasks did not differ for any comparison containing 6 or fewer items (3 vs. 6, 4 vs. 8, 5 vs. 10, 6 vs. 12; all \( p > .1 \)), but did differ for 13/16 of the larger comparisons, (all \( p < .05 \); the trials on which there was no difference were: 24 vs. 48, \( p = .052 \); 100 vs. 200, \( p = .23 \); 300 vs. 600, \( p = .24 \)). This provides some evidence that 5-year-olds have strong associative mappings for numbers up to about 6. For 7-year-olds, the pattern of performance was very similar. They showed no difference in accuracy on the Number Matching task relative to the Discrimination task for the four smallest comparisons tested (all \( p > .05 \)), but they performed significantly worse on the Number Matching task than on the Discrimination task on 13/16 of the larger comparisons (all \( p < .05 \) except for: 24 vs. 48, \( p = .091 \); 60 vs. 120, \( p = .078 \); 100 vs. 200, \( p = .28 \)). Taken together, these data suggest that 5- and 7-year-olds have strong AMs for numbers up to around 6, and have much weaker AMs for larger numbers.

An alternative explanation of these data is that children have AMs for many numbers, but that these AMs are highly inaccurate for all but the smallest numbers. This would cause children to perform poorly on the Number Matching task, especially on large numbers, where a bias to underestimate has been found in both children and adults (Izard & Dehaene, 2008; Siegler & Opfer, 2003). In past studies, such alternative explanations have been ruled out for adults (Sullivan & Barner, 2012). However, to test for this possibility in children, we asked whether subjects showed evidence of over- or under-estimation during the task. If children underestimated the quantities presented, then
they should have performed better on trials where they saw 20 vs. 40 dots and were asked to find forty (foil magnitude is smaller) than those where they are asked to find twenty (foil magnitude is larger). This is because, if the participant erroneously thinks that each set contains fewer items than it actually does, then when forced to find forty, they should strongly prefer the larger quantity (i.e., the correct response, since the smaller quantity will appear even smaller than it actually is, making it a very implausible choice). In contrast, when asked to find twenty, this same participant should often get the trial wrong, and map twenty to the larger of the two sets.

To test this, we compared performance on trials where the foil magnitude was larger than the target to those where it was smaller for each child, in order to determine whether they had an underestimation bias. Next, we classified each child as either showing an underestimation bias or not (there were no overestimators in our population, unlike in adults; Sullivan & Barner, 2012): a child was considered an underestimator if they provided the correct answer significantly more often when the foil magnitude was larger than the target, relative to when it was smaller than the target. Finally, we asked whether our main pattern of findings differed across these two groups. For 5-year-olds, there was a difference in performance between underestimators and those who did not underestimate. While underestimators showed an effect of task and of magnitude (all $p < .0001$), those who did not showed neither effect (all $p > .1$). However, by age 7 there was no such difference: children showed an effect of task (all $p < .0001$), of magnitude (all $p < .0001$), and showed an interaction both if they underestimated ($p < .001$) and if they did not underestimate ($p < .05$). Thus, while 7-year-olds clearly did not possess strong AMs for larger number words, the evidence for 5-year-olds’ use of AM remains
equivocal. To further test the mechanisms guiding children’s number word mappings, we next turn to the estimation task.

**Estimation and Calibrated Estimation**

**Analyses**

Before conducting analyses, we excluded all responses of 0 and 1 (N = 13/3768), as well as all responses more than 10 times larger or smaller than the presented numerosity (N = 538/3768). Additionally, we removed outliers by excluding all data points more than three SD from the mean of each participant’s estimate of each presented set size (N = 36/3768). The frequency of these aberrant responses was comparable in the uncalibrated (14% of data) and calibrated conditions (17% of data), suggesting that they were more likely the result of creativity than of (a) fatigue (which we would expect to increase in the Calibrated Estimation task, since it comes after the Uncalibrated task) or (b) idiosyncratic responding as a result of our calibration manipulation. Analyses were carried out on the remaining 3,181 responses, except for the ordinality analyses, which included all data (because the direction – but not magnitude – of estimates are all that is measured in the ordinality analyses, outliers do not have disproportionate influence on this measure and therefore need not be removed).

We analyzed two measures of estimation performance. The first was the linear relation between the participant’s estimate and the size of the target set. The second was a measure of estimation ordinality. A response was labeled as ordinal if its estimate changed in the correct direction relative to the previous trial. For example, if a larger set was presented on trial \( n \) than on trial \( n-1 \), then the participant’s estimate was considered to be ordinal if it was larger on trial \( n \) than on trial \( n-1 \). In this way, we were able to
measure children’s structural knowledge of the relative ordering of mappings on a trial-to-trial basis, even in cases where estimates were not yet accurate.

Results

First, we conducted a series of analyses to confirm that (a) participants were attending to the task and (b) 7-year-olds outperformed 5-year-olds on the estimation task. Unsurprisingly, a LMM predicting estimates from Magnitude, Age, and their interaction revealed a significant effect of Magnitude ($\beta = .27$, SE = .02, $p < .001$) and a significant interaction of Age and Magnitude ($\beta = .11$, SE = .02, $p < .0001$). Because there was no main effect of Age ($\beta = -5.43$, SE = 4.85, $p > .05$), this analysis suggests that both 5- and 7-year-olds provided estimates that were linearly related to the presented Magnitude, but that their estimates differed for some (but not all) of the magnitudes presented. This is consistent with previous accounts of the development of estimation performance, which have found that 5-year-olds’ estimates of relatively large numbers differ from those of older children, and from their own estimates of relatively small numbers (Siegler & Opfer, 2003; Ebersbach et al., 2008).

In order to test whether children of different ages also differed in their Ordinality score (a measure relevant to assessing SM), we constructed a LMM predicting Ordinality from Age, Magnitude, and their interaction. Again, Magnitude was a significant predictor of Ordinality ($\beta = .002$, SE = .0008, $p < .01$), suggesting that children of all ages were less likely to provide ordinal estimates for relatively large sets. Age was also a significant predictor of Ordinality ($\beta = .316$, SE = .141, $p < .05$), such that older children were more likely to provide ordinal estimates (79% of trials) than younger children (75% of trials). There was no interaction ($\beta = -.002$, SE = .001, $p > .05$). Because of the substantial effect
of Age on estimation performance, we analyzed 5-year-olds’ data separately from 7-year-olds’ data for all subsequent analyses.

In previous studies of adults, changes in estimation behavior caused by misleading feedback have been interpreted to indicate a reliance on SM (Izard & Dehaene, 2008; Shuman, unpublished thesis; Sullivan & Barner, 2012). Thus, our first two analyses tested the effects of calibration to determine whether children recruited SM when making estimates. We also tested whether, as in adults, estimates of relatively small numbers were less influenced by misleading feedback than relatively large numbers.

To answer these questions, we constructed a LMM predicting estimates from the presented magnitude, calibration condition (Calibrated vs. Uncalibrated), and their interaction for each age group separately. Note that these analyses not only test whether feedback influenced estimates for the miscalibrated number (e.g., the largest set estimated), but also for all numbers. For 5-year-olds, we found a significant effect of Magnitude (β = .238, SE = .02, p < .0001), and a significant interaction of magnitude and Calibration type (β = .069, SE = .03, p < .01). For 7-year-olds, we found a similar pattern of results: a significant effect of magnitude (β = .352, SE = .02, p < .0001) and a marginal interaction of magnitude and Calibration type (β = .059, SE = .03, p = .051). This suggests that 5- and 7-year-olds’ estimates were predicted by the target magnitude across all calibration conditions, but that the nature of this relation differed across calibration conditions (see Figure 2.2). Critically, the interaction between Magnitude and Calibration type indicates that estimates of some numerical magnitudes were less affected by calibration than others. This is consistent with the view that some numbers are
mapped via SM – and thus are subject to calibration – while others are mapped via AM and not subject to calibration.

Figure 2.2: Estimates in log-log space by Calibration type for (A) 5-year-olds and (B) 7-year-olds. Data points are individual estimates; lines are model best fits. Calibration condition is indicated next to each best fit line. Note that the degree of separation between calibration conditions increases with magnitude for both 5- and 7-year-olds, even though (because the plots are in log-log space) the separation between conditions appears constant for 5-year-olds.
Models – like the ones reported above – that test the effect of calibration across all magnitudes and all calibration conditions are the clearest way to detect overall the effects of SM. This is because these models allow us to measure the effect of calibration at many different magnitudes, and allow us to compare performance across calibration conditions. Still, one might wonder at which point in the count list Calibrated estimates differed from Uncalibrated estimates. While this type of analysis does not allow us to make predictions about the direction of calibration (e.g., if children underestimate throughout the Uncalibrated task, then even participants in the Calibrated to 25 and Calibrated to 75 conditions might provide larger estimates when Calibrated than Uncalibrated), it allows us to further probe the interaction of Calibration and Magnitude found above. The relatively small amount of data gathered for each set size precludes a detailed analysis of the effect of calibration for each set size, for each calibration type, at each age. However, 5-year-olds showed mean differences (with non-overlapping standard errors) in estimation performance between the Calibrated and Uncalibrated conditions for sets starting as small as 12-32, depending on calibration condition, while 7-year-olds showed effects for sets as small as 6-10 (see Table 2.1). Thus, the effect of calibration, while at times small, emerged across the entire number line, and not just for the very largest numbers tested.
Table 2.1: The smallest magnitudes at which there was an effect of calibration. Note that each Calibration condition represents an independent population of participants. Because of this, direct comparisons of Uncalibrated estimates across Calibration conditions are not meaningful.

Effect of Calibration

<table>
<thead>
<tr>
<th>Set</th>
<th>Mean</th>
<th>St. Dev</th>
<th>Mean</th>
<th>St. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibrated to 25</td>
<td>32</td>
<td>11.5</td>
<td>(1.96)</td>
<td>25.8</td>
</tr>
<tr>
<td>5-year-olds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibrated to 75</td>
<td>12</td>
<td>29.1</td>
<td>(2.7 )</td>
<td>13.1</td>
</tr>
<tr>
<td>Calibrated to 750</td>
<td>16</td>
<td>15.5</td>
<td>(4.0 )</td>
<td>25.4</td>
</tr>
<tr>
<td>Calibrated to 25</td>
<td>10</td>
<td>6.2</td>
<td>(.66 )</td>
<td>17.6</td>
</tr>
<tr>
<td>7-year-olds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibrated to 75</td>
<td>8</td>
<td>6.5</td>
<td>(.87 )</td>
<td>12.4</td>
</tr>
<tr>
<td>Calibrated to 750</td>
<td>6</td>
<td>5.0</td>
<td>(.73 )</td>
<td>13.5</td>
</tr>
</tbody>
</table>

To explore the influence of calibration at the subject level, we asked whether each participant’s estimates differed by Calibration type. Overall, 17 out of the 28 5-year-olds (60.7%) who contributed full estimation datasets showed either a main effect or interaction involving Calibration. By chance alone, we would expect 5% of our population to display a significant effect of Calibration, and an additional 5% to display an interaction. So, we would expect an effect or interaction of calibration to emerge in our models for 10% (e.g., 5% + 5%) of participants by chance alone. However, many more participants showed an effect of calibration than could be accounted for by chance (binomial \( p < .01 \); chance = .1). Also, nine 5-year-olds showed a significant interaction of Magnitude and Calibration, suggesting that these participants incorporated misleading feedback into their estimates in the same way that adults do.

For 7-year-olds, 11/23 participants (47.8%) showed an effect or interaction of Calibration, also far more than would be expected by chance alone (binomial \( p < .01 \)). Of these, 8 showed an interaction, indicating that calibration influenced estimation.
differently as a function of magnitude. Specifically, these participants were less
influenced by misleading feedback for smaller sets, and were more influenced for larger
sets. Given the relatively small number of estimates contributed by each participant, such
large effects of calibration at both the group and individual levels suggests that
participants at all ages relied on SM for most magnitudes tested. More importantly, for
the 17 participants who demonstrated a significant interaction of Magnitude and
Calibration, we see evidence that children recruit strong AM for estimates of small
numbers, but rely more on SM for estimates of larger numbers.

Next, we asked whether participants’ ability to provide ordinal estimates was
mediated by calibration condition. If children rely primarily on SM to construct
estimates, then there should be no difference in ordinality (despite there being substantial
differences in the magnitudes of the actual estimates) across calibration conditions.
However, if calibration does not induce a global shift in number word mapping, and
instead causes children to deploy idiosyncratic estimation strategies, we might expect
ordinality to suffer in the calibrated condition. We found three main results. First,
participants in both age groups provided ordinal estimates significantly more often
(binomial \( p < .0001 \)) than would be expected by chance in the Calibration condition (here,
chance was .5, because any given estimate is likely to be in the correct direction relative
to a previous estimate half of the time; Figure 3). Second, there was no difference in the
likelihood of providing an ordinal estimate across calibration conditions (5-year-olds:
mean .750 ordinal when calibrated, .754 uncalibrated, Chi-squared = .08, \( p > .05 \); 7-year-
olds: mean .799 ordinal when calibrated, .778 when uncalibrated, Chi-squared = 1.1, \( p > .05 \)). Finally, the high rate of ordinality was not restricted only to small sets: by age 5,
mean ordinality was significantly above chance for 7/12 comparisons > 50, and was numerically higher than chance for 11/12 comparisons. This suggests that calibration had a global effect on how number words were mapped to the ANS, consistent with the predictions of SM.

**Counting and Verbal Ordering**

*Counting.* Previous reports have linked counting ability to estimation ability (Davidson et al., 2012; Lipton & Spelke, 2005). However, this relation is complex, and sometimes even poor counters demonstrate strong estimation ability (Barth et al., 2009). Counting performance cannot be used to adjudicate between the use of AM and SM to support estimation, since better counting should be correlated both with more item-based experience and better structural knowledge of number. Still, for the purpose of relating this work to the existing literature, we tested whether better counters (a) provided more accurate estimates; (b) provided more ordinal estimates and (c) were more likely to be influenced by calibration. Highest count for 5-year-old averaged 56.9 on the Free Count test (range 8-100, mode = 100); for 7-year-olds the average was 80.8 (range = 28-100, mode = 100). For the Scaffolded counting assessment, highest count averaged 162 for 5-year-olds (range = 10-1000; mode = 249) and 643.4 for 7-year-olds (range = 100-1000, mode = 1000). The discrepancy in performance between the Free Count test and the Scaffolded counting assessment shows that our participants often possessed knowledge of how relatively large number words are related, even in cases where they were unable to produce those words when reciting the count list. For example, consider the 23 participants who couldn’t count to 100 on the Free Count test. Of these, 10 counted to
100 or higher on the Scaffolded counting assessment, and the mean highest Scaffolded count for this subset of participants was 182. Clearly, even children who lack total proficiency with the routine of counting to 100 still know something about very large number words. One conclusion to draw from this finding is that children may be able to produce sensible estimates for magnitudes that are outside of their productive count range, especially if they possess knowledge of a handful of large number words (see Barth et al., 2009, for evidence that this is sometimes the case). Thus, the highest number a child can count to need not be the biggest number they can use to estimate (and in fact, even our Scaffolded task may underestimate participants’ knowledge).

Next, we asked whether counting performance predicted any of our estimation measures. To do this, we considered counting ability to be a continuous predictor of a variety of estimation outcomes. We analyzed each counting task separately. Neither counting measure predicted the linear slope of participants’ Uncalibrated estimation performance (Free Count: $F((1, 45) = .07, p = .79)$; Scaffolded Count: $F((1, 45) = .87, p = .37$). Also, better counting ability did not predict higher rates of ordinal responding (Free Count: $F((1,45) = .51, p = .48)$; Scaffolded Count: $F((1, 45) = .42, p = .52$).

Finally, better counting ability did not predict a participant’s likelihood that a participant showed a significant effect of calibration (Free Count: Chi-squared = 1.90, $p = .17$; Scaffolded Count: 2.35, $p = .13$). Thus, counting was unrelated to our critical measures of estimation performance, unlike in previous studies (Barth et al., 2009; Davidson et al., 2012; Lipton & Spelke, 2005). This is possibly because many of our children were substantially older than those in previous studies (i.e., such correlations may only exist early in the development of estimation abilities), or because children’s highest count was
not indicative of the actual range of numbers they were familiar with. Future studies should investigate this by testing a wide range of children using a single set of tests.

Figure 2.3: Children’s ordinality scores on the estimation task and verbal ordering task.

*Verbal Ordering.* This task assessed children’s knowledge of the relative ordering of number words in the absence of visual cues (e.g., which is more: *twenty* or *forty*?). Overall, participants performed well at this task, and both 5- and 7-year-olds provided correct responses more often than would be expected by chance alone (5-year-old mean = 71.4%; 7-year-old mean = 91.6%; binomial *p*-values < .0001; see Figure 2.3). Children’s accuracy improved substantially with age (*F*(1, 45) = 22.40, *p* < .0001), suggesting that knowledge of verbal ordering increases between the ages of 5 and 7. These data suggest that children in both age groups possess some understanding of the relative ordering of number words, a skill that is required for the adult-like deployment of SM. However, there was no relation between accuracy on the verbal ordering task and the likelihood that a participant would provide an ordinal response on the estimation task (*β* = .01, SE = .07, *p* > .05), and while Verbal Ordering performance improved substantially with age,
Estimation Ordinality improved only a small amount (see Figure 2.3). While possessing ordinal number word representations is a prerequisite for forming an SM, the presence of relatively good verbal ordering does not ensure that estimates will be ordinal.

**Discussion**

When children connect language to ANS representations of number, they rely heavily on inferential processes to do so. Based on data from six tasks, we found converging evidence that children, like adults, recruit both Associative Mapping (AM) and Structure Mapping (SM) to construct estimates, suggesting that inferential and associative processes are fundamental to the formation of number word mappings. However, while we found that children use both mechanisms to support estimation, children possessed fewer strong AMs than adults, and appeared to rely heavily on inferences about the structure of the number system. Taken together, our tasks converge to suggest that – from early in development – structural inference is fundamental to guiding connections between number language and number perception. This finding highlights not only the importance of SM in supporting numerical knowledge, but also provides a window into understanding the mechanisms that guide children’s learning about the relation between language and perception.

According to the AM hypothesis, each number word is mapped onto an ANS representation of numerical quantity on an item-by-item basis, as a result of experience with particular word-magnitude pairings (e.g., Lipton & Spelke, 2005). Converging evidence for children’s use of AM comes from two sources. First, on the Number Matching task, participants tended to be accurate at matching a number word to one of two discriminably different sets when the sets contained a small number of dots:
Specifically, both 5- and 7-year-olds showed no differences in performance on the Discrimination and Number Matching tasks for sets containing six or fewer items, but showed large differences for all larger magnitudes. Second, on the Calibrated Estimation task, many children, like adults (Sullivan & Barner, 2012), were less influenced by misleading feedback for small numbers than for large numbers. Taken together, these findings suggest that our participants possessed strong, statistically reliable, AMs for numbers up to at least six. Of course, it is unlikely that there is anything special about the number 6 as a cutoff between AM and SM, since the strength of AMs appear to gradually decline as a function of magnitude (and thus the apparent cutoffs is only reflected in significance testing, and not in the pattern of effects themselves). In adult populations, there is large individual variability with respect to where individual subjects exhibit significant differences (Sullivan & Barner, 2012), and this would almost certainly be the case in child populations as well given a paradigm that allowed testing individual differences. Our data also do not differentiate between the possibilities that (a) children possess no AMs above six or (b) children possess AMs that are weaker for larger numbers. However, our data do clearly demonstrate that insofar as AM guides estimation, it plays the largest role in supporting estimates for relatively small numbers, and a much smaller role in supporting estimates for large numbers.

In contrast, the SM hypothesis posits that each number word mapping is constructed in relation to all other mappings, and that the verbal and nonverbal number systems become related to each other on the basis of similarities in their structures. Evidence for SM came from three sources. First, children’s accuracy on the Number Matching task declined as a function of numerical magnitude relative to the
Discrimination task (resulting in a significant interaction for seven-year-olds), something that would not be predicted if all number words were mapped with equal strength on an item-by-item basis. Second, children’s estimates were influenced by misleading feedback about the largest set on the Calibrated Estimation task, showing that most number word mappings are mutually constraining: alterations to one mapping influenced many other mappings. This finding – that feedback about the largest set influenced estimates of other sets – has also been demonstrated in adults (Sullivan & Barner, 2012), suggesting that children and adults recruit similar mapping mechanisms. Third, during estimation tasks, children’s responses tended to be ordinal (i.e., in the correct direction relative to previous estimates) regardless of estimation accuracy or calibration condition. Our data showed that even the worst estimators – the 5-year-olds – consistently provided ordinal responses more often than would be expected by chance. Taken together, these data suggest that even very young children make inferences based on the structure of the count list when estimating.

At the outset, we noted several ways in which AM and SM might combine to support estimation. One possibility was that children might rely heavily on AM to construct their first number word mappings between language and the ANS (Lipton & Spelke, 2005), and only begin to recruit SM later in development. Second, we noted that, early on, children might first acquire an adult-like set of AMs (e.g., strong up to around 12) before reliably using SM to guide estimation. A final possibility is that children might initially rely heavily on SM, taking advantage of structural knowledge of the count list to form an inferentially-derived set of mappings, before acquiring most AMs. Of these three alternatives, our data are most consistent with the third: both 5- and 7-year-olds were
relatively unaffected by calibration for small numbers up to around 6, and performed well for these numbers on the Number Matching task. However, our data are also potentially consistent with a version of the first hypothesis. It is possible that children initially have no AMs between number words and the ANS, but instead that AMs are mediated by other types of numerical representations. This view is possible under the hypothesis, controversial in some quarters, that estimation for the smallest quantities – e.g., 1 to 4 – is supported by “parallel individuation” – i.e., a system for tracking multiple objects as they move through space (see Carey, 2009; Feigenson et al., 2004, for review; for details regarding multiple object tracking in adults, see Pylyshyn & Storm, 1988). On this view, children who estimated using parallel individuation, rather than the ANS, might have made robust estimates for numbers up to 6 by quickly subitizing a subset of an array, and then accurately extrapolating this estimate to the larger set (e.g., using an additive or multiplicative function; though see Cordes & Brannon, 2009; Negen & Sarnecka, 2010; vanMarle & Wynn, 2009 for evidence that the ANS could be used for small sets). In the context of the present study, it is impossible to differentiate this possibility from the idea that estimation for all numbers involves the use of the ANS. More importantly, if AMs between language and the ANS exist at all, they are very limited in scope, and appear to be strongest for small, frequent, and thus familiar number words (for a discussion of the relative frequency of small vs. large number words, see Dehaene & Mehler, 1992). Thus, to the extent that associative mappings exist in children, they play a minimal role, whereas even our youngest participants showed evidence of SM in their estimation behavior, supporting the view that inference and analogy are essential learning mechanisms in forming number word mappings.
Importantly, 7-year-olds did not appear to have stronger associative mappings than 5-year-olds. Thus, children’s AMs (however they are represented) did not change noticeably over a period of development in which significant improvements in estimation accuracy occur. This suggests that developmental improvements in estimation (e.g., Siegler & Opfer, 2003; Siegler & Booth, 2004) are not due to changes in associative mappings. Relatedly, while some have hypothesized that children must have a core set of associative mappings in order to support SM (Carey, 2009; Sullivan & Barner, 2012), these data suggest that this core set is very small. Children were able to recruit SM even when they only possessed robust AMs up to about six. This finding strongly suggests that the process of learning to estimate is one that is guided primarily by structural inference.

This finding – that AM plays a limited role in the development of estimation abilities, whereas SM plays a larger role – is relevant to several important questions previously debated in the estimation literature. First, this study is relevant to the observation that estimation ability in school-aged children predicts success in mathematics (e.g., Booth & Siegler, 2008; Siegler & Ramani, 2009; Siegler & Booth, 2004). From our data, it appears unlikely that those who are better at estimating (and thus better at math) have a relatively richer set of associative mappings. While educators often focus on providing manipulatives (e.g., toy blocks) to help children visualize the quantities symbolized in math problems (e.g., Burns, 1996; see Uttal, Scudder & DeLoache, 1997, for another view on the role of manipulatives), even our strongest estimators lacked strong AMs, suggesting that item-specific connections between language and visual representations of magnitudes may not drive early math success.
A different explanation of the relation between estimation and education outcomes is that both draw on children’s abilities to recruit SM. Our data showed no relation between counting and estimation ability, and this suggests that merely learning the routine of counting does not ensure adult-like knowledge of the structural relation between numbers within the count list. However, our findings do not rule out the possibility that math skill and estimation ability are related via their shared reliance on SM. Early math education often focuses heavily on teaching the structure of the verbal number system. For example, explicit instruction about the place value system reinforces structural information about the relations between number words—by understanding place value, children might learn the relation between, say, 30 and 300 (or, conversely, a strong understanding of the relation between 30 and 300 might make it easier for some children to learn concepts like place value). Similarly, basic arithmetic processes involve relating symbolic representations of number to each other. Children who know that $20 + 20 = 40$ may be better estimators because both estimation and early arithmetic draw on knowledge of the relation between number words. If this is the case, then focused instruction on the structure of the count list (and not just the routine of counting) may be the best way to improve both math and estimation outcomes. Future research will be required to explicitly test this claim, and to investigate the possible relation between SM and math success.

In this paper, we have demonstrated how associative and inferential processes interact to guide children’s estimates during development. We have shown that children as young as 5 recruit both AM and SM when estimating, and that they do so in remarkably similar ways to older children and adults. Both 5- and 7-year-olds possess
strong associative mappings for numbers up to about *six*, and both recruit SM for larger numbers. Future work is required to model the particular ways in which these two learning mechanisms interact throughout development and into adulthood, with the specific goal of understanding the learning mechanisms guiding developmental changes in estimation. Characterizing these changes will help us to understand language and perception interact, and how inference and item-specific experiences combine to form shape conceptual knowledge of number.
Chapter 2, in full, is a reprint of the material as it appears in Inference and Association in Children’s Early Numerical Estimation. *Child Development*, doi: http://dx.doi.org/10.1111/cdev.12211. Sullivan J. & Barner, D. (in press). The dissertation author was the primary investigator and author of this paper. Permissions for use of this material have been obtained from John Wiley and Sons, Ltd.
References


Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition, 86*, 201-221. DOI: 10.1016/S00100277(02)00178-6


Huntley-Fenner, G. (2001). Children’s understanding of number is similar to adults’ and rats’: numerical estimation by 5-7 year olds. *Cognition, 78*, B27-B40. DOI: 10.1016/S0010-0277(00)00122-0


psychophysics of number representation. *Psychological Science, 10*, 130-137. DOI: 10.1111/1467-9280.00120


Chapter 3

The development of structural analogy in number-line estimation
Abstract

Recent studies have revealed that making number-line estimates not only requires number knowledge, but also a host of other cognitive skills (Anobile, Stievano, & Burr, 2013; Barth & Paladino, 2011; Kolkman, Hoijtink, Kroesbergen, & Leserman, 2013; Slusser, Santiago, & Barth, 2013). Here, we argue that a fundamental component of number-line estimation is the act of relating the target number being estimated to another numerical reference point (e.g., a previous estimate; the endpoint of the line) and then extending this relation to the spatial domain – in other words, that children recruit analogical reasoning skills when estimating. Because such analogical comparisons require both the selection of a numerical reference point and the comparison of that reference point to the target number, we aimed to understand which reference points children use and how they use them. To this end, we tested whether and how 5-, 6-, and 7-year-old children used their previous estimates to constrain subsequent estimates. We found that children used their previous estimates as reference points, that older children used reference points differently than did younger children, and that the ability to access previous estimates limited our youngest participants’ ability to perform well on our number-line estimation task. We conclude that the analogical reasoning component of number-line estimation is substantial, and shapes children’s earliest estimation performance.

Keywords: Estimation, numerical cognition, analogy
Number-line estimation – i.e., the ability to place numbers on a physical line in correspondence to their relative magnitude – is often used to test theories of number knowledge, math skill, and cognitive development more broadly (Booth & Siegler, 2006; Ebersbach et al., 2008; Moeller, Pixner, Kaufmann, & Nuerk, 2009; Siegler & Booth, 2004; Siegler & Opfer, 2003; Siegler & Ramani, 2008). Estimation performance is predictive of math achievement (Booth & Siegler, 2008; Ramani & Siegler, 2008; Schneider, Roland, & Paetsch, 2009; Siegler & Ramani, 2008) and memory for numbers (Thompson & Siegler, 2010), suggesting that the skills and knowledge measured by number-line estimation tasks are important to general academic success. While most experiments on number-line estimation focus on how changes in estimation ability relate to the development of number knowledge, some recent studies have shown that number-line estimation depends not only on number knowledge, but also on a cluster of additional capacities not directly related to number (Anobile et al., 2013; Barth & Paladino, 2011; Kolkman, et al., 2013; Slusser, Santiago, & Barth, 2013). As a result, developmental changes in number-line estimation (and their correlations with other important cognitive factors) may stem in large part from changes to non-numerical capacities.

Many studies have found that children’s ability to accurately place numbers on a number line improves gradually over development, and that this improvement is characterized by a shift towards accurate, linear estimation behavior within the child’s familiar number range (e.g., Booth & Siegler, 2006; Siegler & Opfer, 2003; Slusser et al., 2013). These changes to estimation performance depend, in part, on changes to children’s understanding of the verbal number system. As children gain familiarity with a particular
numerical range (e.g., 0-100), estimation improves in that range (but not, necessarily, for unfamiliar numbers – e.g., 101-1000). However, as noted above, improvements to estimation behavior may not indicate improvements to number knowledge alone: recent work has also related the development of number-line estimation ability to constructs as diverse as executive function and spatial proportion reasoning (Barth & Paladino, 2011; Kolkman et al., 2013; Slusser et al., 2013). These studies raise the possibility that improvements to estimation behavior arise not only from improvements to number knowledge, but also from improvements to more domain-general cognitive skills. These studies also raise the more general point that estimation tasks may not just measure the accuracy of children’s mental number line, but may also measure other types of knowledge and skills.

In the present study, we explored the role of one such ability – analogical reasoning – in supporting number-line estimation. How might estimation ability require analogical reasoning skills? In order to understand the role of analogical reasoning in estimation performance, it is necessary to understand the processes involved in analogy, and how these might apply to number-line estimation (for discussion, see Carey, 2009; Cantlon, Cordes, Libertus, & Brannon, 2009; Sullivan & Barner, 2012). Very generally, an analogy involves picking out a relation between entities in one domain and then applying this relation to entities in a separate domain. For example, consider the analogy: “a mitten is to a hand as a sock is to a foot” (or, in standard analogical notation, Mitten : Hand :: Sock : Foot). This analogy involves two components. The first is a within-domain comparison (e.g., Mitten : Hand), which is used to extract the relation(s) between the entities in one domain (e.g., that a mitten goes on a hand, or that mittens keep hands
warm). The second is an across-domain comparison, which applies the relation picked
out by the within-domain comparison to entities in the new domain (e.g., that a sock goes
on a foot, or that socks keep feet warm).

Number-line estimation tasks involve precisely this kind of analogy. Because both
the length of number lines and the range of values that they represent can vary from one
line to the next, the only way to construct an estimate is to relate the number being
estimated to other numerically meaningful reference points. In other words, on any
particular trial, the appropriate location for an estimate can only be determined by first
comparing the number being estimated to other numerical reference points on the number
line, and by then translating the resulting numerical relation into an analogous spatial
representation. For example, consider an estimate of ‘50’ that is located 2” from the 0
point on a number line. Determining whether this estimate is accurate is impossible
without knowing the location of another numerical reference point, like the numerical
upper bound of the line (e.g., 100 vs. 1,000,000). So, as in the mitten : hand analogy
described above, the ability to make number-line estimates depends on establishing a
within-domain relation between items in the number list, and extending this relation to
spatial points on the number line (see Barth & Paladino, 2011 for a discussion of the
spatial component of this analogy).

In this paper, we sought to understand how analogical processes contribute to
changes in estimation performance. As noted above, when children are asked to estimate
a number, they must make a within-domain numerical comparison between the number
being estimated and other numerical reference points on the number line. This involves at
least two steps: (1) the child must first select a numerical reference point to be used as a
point of comparison, and (2) they must assess the relation between that reference point and the number being estimated. Only once a child has taken these two steps can they complete the analogy by representing the relation between the two numbers spatially on the number line. Thus, in order to understand the role of analogical reasoning in number-line estimation, it is crucial to understand both (1) which reference points children select when estimating, and (2) which numerical relations they pick out between these reference points.

With respect to reference point selection, children have two main options. One option is to use reference points that are provided to them directly by the experimenter. For example, when estimating ‘50’ on a 0-100 number line, children might select ‘100’ as an appropriate reference point because it is the endpoint of the number line. To complete the within-domain numerical comparison, they then use this reference point to calibrate their estimate, by picking out some relation (described in detail below) between ‘100’ and ‘50’. Past studies indicate that older children and adults use reference points like the end-point and mid-point of the number line to calibrate their estimates (Barth & Paladino, 2011; Siegler & Opfer, 2003; Sullivan, Juhasz, Slattery, & Barth, 2011; for related findings in dot array estimation tasks, see Izard & Dehaene, 2008; Sullivan & Barner, 2012; in press). However, in some cases it may not be possible to use midpoints and endpoints as reference points, since using these experimenter-provided points often requires knowledge of relatively large numbers (like 100) that some young children lack (Barth & Paladino, 2011).

Alternatively, in an experiment with multiple trials, children could make use of their own previous estimates – e.g., by recalling their previous response of 10 during a
trial in which they are asked to make an estimate for 50 (Sullivan et al., 2011; Vul, Sullivan, & Barner, 2013). This strategy might allow children with limited numerical knowledge to use reference points from their familiar number range. However, this strategy is also constrained, in this case by the child’s ability to access a previous estimate and use it to guide a numerical comparison. Since number-line experiments typically ask children to make one estimate per number line, children’s estimates from previous lines can often be accessed only via memory. If children use their previous estimates to calibrate their subsequent estimates, then the ability to encode, recall, or mentally manipulate the numerical magnitude – or spatial location – of their previous estimate may limit their ability to make accurate estimates. Consequently, if children are unable to use experimenter-provided reference points (e.g., because these numbers are too large), and are unable to access responses from previous trials, then they may fail to make accurate estimates despite having strong internal representations of how numerals encode number. Currently, it is unknown whether young children use their previous estimates to construct new number-line estimates, and to what degree their early problems with number-line estimation are related to domain-general cognitive difficulties in accessing previous responses when calibrating estimates. In this paper, we explore the question of which reference points children use when making number-line estimates.

A separate question that we explore in this paper is which numerical relations (between reference points and the number being estimated) children make use of when making estimates. As noted earlier, analogies can encode a number of different relations between entities. For example, in the “hand : glove” example, a child might focus on the fact that a hand goes inside a glove, or, alternatively, that gloves keep hands warm.
Similarly, in the case of estimation, individual children may encode different relations between numerals, with consequences for their ability to make accurate, adult-like estimates. Here, we focus on two possible relations: ordinality and relative distance.

The first within-domain relation that children might construct is an ordering between the number being estimated and the numerical reference point. For example, a child who relies on ordinality to estimate the location of ‘50’ (using the endpoint of the line ‘100’ as a reference point) should first note that ‘50’ is smaller than ‘100’ and therefore place his estimate somewhere to the left of the endpoint. A different child who also recruits knowledge of ordinality, but uses her previous estimate (e.g., of ‘10’) as a reference point will place her estimate of ‘50’ somewhere to the right of where she had placed ‘10’, because ‘50’ is larger than ‘10’. Using the ordinal relation between the number being estimated and the selected numerical reference point to calibrate estimates is an unsophisticated strategy: it does not guarantee accurate responding. However, it also does not require knowledge about the relative distances between numbers in the count list, making it a potentially viable strategy for estimating, especially for young children who do not yet have sufficient mastery of the number system to accurately construct more sophisticated relations (discussed below).

The second within-domain relation that children might construct is the relative distance between the target number being estimated and a numerical reference point. For example, a child who has knowledge of the relative distance between numbers might note that ‘50’ comes halfway between the beginning of the count list and ‘100’, and use this information to place their estimate at the midpoint of the line. A child who uses their previous estimate – e.g., of ‘10’ – might note that ‘50’ refers to a quantity that is 5 times
greater than is referred to by ‘10’, and therefore place ‘50’ 5 times further from the start of the number line than they had placed ‘10’. Although this strategy is more advanced than ordinality alone (since it involves both placing numbers in order and representing their relative distances), it also does not guarantee accurate responding, especially if children begin from a reference point that is not accurately situated on the number line (e.g., a previous estimate).

To investigate the role of analogy in number-line estimation, we tested how children select and reason about numerical reference points when estimating. To do so, we conducted an experiment that differed from previous studies in three critical ways. First, we asked whether children attended to the numerical ordering of new estimates relative to their previous estimates. Critically, it is possible to provide inaccurate responses while still providing ordinal responses. However, analyses typically used to assess estimation are not designed to detect ordinal responding in the absence of accuracy. Ours is the first study to test children’s use of previous estimates as reference points during number-line estimation tasks, and whether children respect the ordering of these estimates (see Sullivan & Barner, in press for some evidence of this strategy in array estimation tasks). Specifically, we asked whether individual estimates were ordinal relative to previous estimates (independent of their accuracy). Doing this allowed us to assess the extent to which children – at all ages – made analogical comparisons between the number being estimated and previous estimates.

Second, we asked whether children attended to the relative distance between new estimates and their previous estimates. We tested this in two ways. First, we simulated estimation data for “participants” who relied only on the relative ordering of adjacent
numbers when generating estimates. We then asked whether the simulated data was similar to our participants’ actual estimates (consistent with reliance on trial-to-trial ordinality alone), or whether their estimates were better explained by the use of both ordinality and relative distance information. Second, we introduced a “distribution manipulation” through which some children were asked to make estimates of number drawn from a distribution of relatively smaller numbers, while others were asked to make estimates from a distribution of relatively larger numbers. This manipulation was designed to test children’s use of relative distance information when relating previous estimates to later ones. As reported by Sullivan et al., (2011), participants tend to underestimate large numbers and overestimate small numbers. When overestimation or underestimation happens early in a task, participants who use previous estimates to calibrate later ones will carry this error forward throughout the remainder of the task. However, early differences in estimation should not carry through to later trials if (a) children exclusively use stable reference points on the number line to calibrate their estimates or (b) children use previous estimates, but only to represent the ordering of numbers and not the relative distance between them. Thus, this manipulation allowed us to test whether children represented the relative distance between early estimates and later ones, by measuring whether their estimates – across the entire task – differed across Distribution conditions.

Third, we reasoned that, as in other cases of quantity comparison (e.g., Bryan & Trabasso, 1971), if children make use of previous estimates to constrain subsequent ones (by comparing the target number being estimated to previous estimates), their estimation performance should depend on the their ability to access previous responses. This view of
estimation is especially plausible in light of recent research that has shown that children’s ability to remember and update mental representations on-line is a strong predictor of estimation performance (Kolkman et al., 2013). Typical estimation tasks require children to make estimates one at a time. If children don’t use their previous estimates as reference points, then this property of the number-line estimation task should be irrelevant. However, if children do use their previous estimates as reference points, then not having access to previous estimates could limit children’s performance – they may have difficulty remembering or accessing the numerical magnitudes that they previously estimated (thus limiting their ability to make a within-domain numerical comparison), and they may have difficulty remembering or accessing the location of the previous estimate (thus limiting their ability to analogically extend their numerical comparison to the spatial domain). To test whether having visual access to previous estimates influences estimation performance, we provided some participants with visual access to their previous estimates, and compared their estimation behavior to children who received a standard number-line estimation task, where they were asked to make estimates one-at-a-time. This methodological choice built on previous studies that have asked children to place multiple estimates on a single line (Siegler & Booth, 2004). By making several changes to the previous paradigm (e.g., testing our participants on a wide range of numbers, not providing a training period, and considering measures other than accuracy), we were able to leverage this method to test the role of analogical reasoning in early estimation.¹⁰ Thus, this is the first study to test the role of analogy in children’s early

¹⁰ Siegler and Booth (2004), only tested 5-year-olds with very small numbers (0-10), whereas we tested 1-100, which allowed us to test the use of analogy for numbers outside
estimation ability by considering developmental changes in (1) which reference points
children use; and (2) which relations children pick out between estimate and reference
point.

EXPERIMENT

Materials and Methods

Participants

Eighty-five children participated. Seventy-seven children completed at least 24
trials and were included in the final analyses. This included 26 5-year-olds, 25 6-year-
olds, and 26 7-year-olds (range 5;0-7;11).

Materials

Stimuli consisted of horizontal black number lines that were 23 cm long. Individual
lines were centered on paper measuring 4.25” x 11”. Printed on the left of each
number line was the numeral “0” and on the right was the numeral “100”. The numbers
to be estimated were presented auditorily, and ranged from 3-97. There were no visual
cues as to the location of the midpoint of the line (Barth & Paladino, 2011).

Procedure

Participants were shown the number line and were told, “This is a number line.
See? It goes from 0 all the way to 100”, while the experimenter gestured from left to right
across the length of the line. The experimenter continued, “Each number has its own
special place on the number line. Today, you’re going to show me where certain numbers
go on the number line. Look! Zero goes here [gesture to leftmost endpoint] and 100 goes
here [gesture to the rightmost endpoint]. And all of the other numbers have their own
special places on the number line. I’m going to give you a pencil, and your job will be to
draw an up-and-down line to show me where each number goes. Are you ready?”
Participants were then given 24 estimation trials. On each trial, the number to be
estimated was provided, and the child was given a new, differently colored pencil to mark
each answer (to differentiate estimates when they were marked on the same sheet). The
task took approximately twenty minutes to complete.

In order to test whether the accessibility of previous estimates influenced
children’s estimates, participants were randomly assigned to one of two conditions: a
Single Estimate condition or a Multiple Estimate condition. In the Single Estimate
condition, as with previous studies of children’s number-line estimation, participants
made estimates for numbers one at a time, marking each estimate on a new number line
(see Booth & Siegler, 2006; Siegler & Opfer, 2003). In the Multiple Estimate condition,
participants made estimates one at a time, but provided multiple estimates on the same
number line (e.g., Siegler & Booth, 2004). To avoid cluttering the number line, estimates
for the first 12 trials were recorded on one number line, and the last 12 trials were
recorded on a separate line (only one number line was visible at a time). Twelve children
(14% of our test population) asked to be reminded of a previous estimate (e.g., “What

11 Approximately 60 of the participants were given the opportunity to complete a
second set of 24 trials in the opposite condition. Due to significantly higher rates of error
across all age groups, and numerous experimenter notes of inattention during the second
24 trials, data for these trials were not analyzed further.
number was the pink line?” or “Can I see the other sheet?”) – their requests were honored in the Multiple Estimate condition ($n = 5$) but were denied in the Single Estimate condition ($n = 7$).

Participants estimated numbers drawn from one of two possible distributions. In the Small Number Distribution, 24 numbers were selected between 1-100 such that 4 were smaller than 10: 3, 4, 6, 8, 12, 14, 17, 18, 21, 24, 25, 29, 33, 39, 42, 48, 52, 64, 72, 79, 81, 84, 90, 96 (Booth & Siegler, 2006). The Large Number Distribution contained the 24 numbers generated by subtracting the Small Number set from 100 (Barth & Paladino, 2011; Sullivan, et al., 2011). There were two possible trial orders within each Distribution. These were pseudo-random permutations of the selected numbers, arranged so that the first few trials contained a small number (for those in the Small Number Distribution set) or a large number (for those in the Large Number Distribution set).

Children’s estimation behavior was also qualitatively coded online for evidence of reference-point use, counting, and other strategies. Preliminary analyses indicated that these measures were not related to any of the data reported here, and thus they are not included.

**Simulation**

We conducted several simulations of estimation data in order to formalize our predictions about the empirical effects of (a) using previous estimates to calibrate later ones and (b) our Distribution manipulation. Simulations were conducted using custom MatLab code or in Excel (both available upon request), and were constrained by theoretical accounts of estimation performance (described below). Each simulated experiment contained 24 “participants” randomly assigned to one of our two Distribution
conditions (and also to one of our two trial orders). Each simulation was iterated 1000 times.

**Trial-to-Trial Ordinality.** This simulation randomly generated an estimate between 0-100 for the first trial, and then followed trial-to-trial ordinality $x\%$ of the time (where $x$ was the percentage of ordinal responses at each age group in our real data, see Results). For example, if on Trial 1 an estimate was randomly selected as ‘25’, then on Trial 2 the response for a large number would fall between 25.01 and 100, whereas the response for a smaller number would fall between 0 and 24.99, thereby respecting ordinality in each case. We ran 1000 iterations of this simulation for each simulated age group (5 YOs; 6 YOs; 7 YOs).

**Trial-to-Trial Relative Distance.** This simulation generated a first estimate by taking the target number being estimated (e.g., ‘18’) and randomly selecting an estimate that was +/- $[0-40\%]$ of the target number (e.g., 18 +/- [0-7.2]). We selected the range 0-40% error because that is most likely to converge at a PAE of 20%, which is what we found in our real data (see Fig. 3.1).

\[12 \text{ Note that if the number requested on the first trial is relatively large, this means that the model is probabilistically likely to underestimate that first number. Also note that after the first trial, an estimate of a large number is probabilistically likely to be followed by a request for a smaller number. In this way, estimating from a Large Number distribution can lead to systematic underestimation if participants use previous estimates to calibrate later ones.}\]

\[13 \text{ We also simulated other amounts of error (e.g., 0-10\% error) for this and all other simulations in which error rates were specified in the simulation. Results were quantitatively and qualitatively similar to those reported here. When testing for an Interaction of Distribution and Magnitude (see below), we found that as error decreases, the Trial-to-Trial Relative Distance strategy is even \textit{more} likely to yield an Interaction, while all other strategies are \textit{less} likely to.}\]
To generate the 2nd estimate, this simulation calculated the ratio of the target number on Trial 2 to the target number on Trial 1 (e.g., if Trial 2 was ‘36’, the ratio would be 2:1), and then multiplied the 1st estimate by that ratio. In this way, the relative distance between estimates was preserved, regardless of the accuracy of the initial estimate, and each subsequent estimate was contingent on the immediately preceding estimate.

**Independently Generated Estimates.** We ran two simulations in which estimates were generated independently from one another (e.g., without referencing previous estimates). For the first simulation, on each trial the estimate was a randomly selected number that was +/- [0-40%] of the target number.

The second simulation took into account participants’ bias to overestimate small numbers and underestimate large ones. If the number being estimated was larger than 50, the program generated an estimate that was 0-40% below the number being estimated; if it was smaller than 50, the program generated an estimate that was 0-40% above the number being estimated.

**Analyses**

Children’s responses were converted to their numerical estimate equivalent. Indecipherable responses were excluded (n=9/1848 trials). Responses that were located immediately to the right of the number line’s endpoint were included in the final analyses (n=28/1848 trials) as these were frequently accompanied by a child’s explanation (e.g., “This one has to be off the list”). These responses resulted in some estimates that were larger than 100 (see also Cohen & Blanc-Goldhammer, 2011, for a discussion of how the bounds of a number line can constrain estimates in undesirable ways, and why the
assessment of numerical knowledge can be facilitated by using unbounded number-line tasks). Analyses excluding these 28 trials were also conducted, with identical effects to those reported below.

Our analyses focused on two measures of estimation performance. First, we used regressions to analyze the relation between each estimate and the number being estimated (e.g., Barth, Starr, & Sullivan, 2009; Booth & Siegler, 2006; Lipton & Spelke, 2005; Siegler & Opfer, 2003). This measure is often used to quantify the accuracy of estimates, because it allows researchers to test the relation between the number being estimated and the child’s estimate, with slopes closer to 1 typically indicating more accurate performance. However, this analysis does not test whether inaccurate responses are ordinal or random. To this end, we also measured whether the child’s estimates respected the ordinality of the count list. A trial was labeled as ordinal if the child provided an estimate in the correct direction relative to a previous estimate, regardless of its accuracy (e.g., by providing a larger estimate on trial n than on trial n-1 if and only if a larger number was requested on trial n than on trial n-1). Our main ordinality test asked whether any given estimate was ordinal relative to the immediately preceding estimate (thus comparing trial n to n-1).

With the exception of binomial comparisons to chance and chi-squared analyses, all analyses reported below were conducted using the LME4 package of R (Bates & Sarkar, 2007; R Development Core Team, 2010). All models were Linear Mixed Models, with Subject considered a random factor. Condition (Single Estimate vs. Multiple Estimate) and Distribution (Small Number Distribution vs. Large Number Distribution) were considered fixed factors. Ordinality scores resulted in binomial data, and were
therefore subjected to binomial logit analyses. Simulated data and real data were analyzed using the same code. We report parameter estimates (β), MCMC p-values estimates, and standard error estimates. All results reported below are of real (child-generated) data unless an analysis is described as being based in our computer simulated data.

**Results**

*Estimation Performance and Replication of Past Data*

We predicted participants’ estimates from a model containing Age and the Magnitude of the number being estimated. Consistent with previous research, there was an effect of Age such that older children provided different estimates than younger children (β = 11.5, SE = 1.6, p < .0001), an effect of Magnitude indicating that estimates were related to the number being estimated (β = .66, SE = .11, p < .0001), and an interaction of Magnitude and Age, indicating age related differences in the specific relation between the target number and the child’s estimate (β = -.20, SE = .02, p < .0001).

Next, we predicted participants’ estimates from the numerical Magnitude they were estimating, separating participants by Age. Here, β can be interpreted as a simple slope measure, where the closer β is to 1 the closer estimates are to what we called “adult-like” performance (although even children with slopes of 1 may differ in important ways from adults). Predictably, 5-year-olds performed the worst, although they still made estimates that were linearly related to the target Magnitude (5-year-olds: β = .36, SE = .03, p < .0001). Six-year-olds’ estimates had a slope closer to 1, indicating more adult-like performance (β = .57, SE = .02, p < .0001), and 7-year-olds performed extremely
well ($\beta = .74$, $SE = .02$, $p < .0001$). Thus, younger children’s estimates were somewhat inaccurate, and did not display an adult-like linear relation between number and estimates.\textsuperscript{14}

Finally, we asked whether percent absolute error in estimation (or PAE) increased with the magnitude of the target number, where PAE was defined by the absolute difference between the number being estimated and the child’s estimate. Generally, when estimation tasks draw on the Approximate Number System (ANS), estimation error increases with number (Dehaene, 1997; Whalen, Gallistel, & Gelman, 1999). However, past studies have found that number-line estimates do not show this signature of the ANS (e.g., Siegler & Opfer, 2003), and thus that estimates which depend on analogical numerical comparisons are unlikely to be linked to the ANS. We tested whether the PAE increased with number (indicating that estimates were supported by the ANS). Consistent with past reports, we found that it did not ($\beta = -.007$, $SE = .013$, $p > .5$; see Fig. 3.1), suggesting that the ANS did not support participants’ number-line estimates.

\textsuperscript{14} To ensure that our data replicated recent reports of children’s estimation, we also assessed the degree to which proportional reasoning models of estimation could account for deviations from linear performance found in our data. We found that, as reported previously (Barth & Paladino, 2011; Cohen & Blanc-Goldhammer, 2011; Sullivan et al., 2011), deviations from linearity could be partially accounted for by the error introduced in proportional judgments. We also tested logarithmic fits, and found that our youngest participants tended to be fit well by these curves. Because our interest was not in the particular shape of the estimation function involved, but rather in children’s use of analogy and working memory, we do not discuss these models further.
Use of Previous Estimates and Reliance on Ordinality Relations

Critical to the present proposal – that developmental changes to estimation behavior may be linked to changes in the types of reference points used during estimation tasks – we tested (a) whether children compared their earlier estimates to later estimates and (b) how children used those reference points. To this end, we analyzed children’s ordinality scores, which represented whether they preserved numerical ordering relations across adjacent trials. As described above, for each child, we computed whether a particular trial’s estimate was ordinal relative to the immediately preceding estimate. So, if a larger number was requested on trial $n$ than on trial $n-1$ a child’s response was counted as ordinal if and only if it was larger for trial $n$ than for trial $n-1$. Because it is
impossible to compute ordinality for the first trial of the task (since there is no preceding trial), this trial was not scored.

Our logic was that children who provide inaccurate estimates could be doing it for one of two reasons. First, they could be performing randomly (e.g., out of lack of numerical knowledge or confusion). Alternatively, poor estimators may lack knowledge of the relative distance between numbers, but may still represent the ordering of the target number relative to their previous estimate. If children represent the relative ordering of number words and use this to guide their number-line mappings, then they should demonstrate above-chance levels of ordinality.

First, we asked whether children provided ordinal responses. In a binomial logit regression, we predicted the ordinality of estimates from Age and found that the likelihood of ordinal responses increased significantly as a function of Age ($\beta = 1.02$, $SE = .16$, $p < .0001$). Although young children’s estimates were less likely to be ordinal than the estimates of older children, all participants demonstrated high levels of ordinality: Five-year-olds provided ordinal responses on 69.6% of all trials, 6-year-olds did so on 84.1% of trials, and 7-year-olds did so on 93.3% of trials (see Fig. 3.2). There was a significantly greater proportion of ordinal trials within each age group than would be expected by chance alone (chance = .5, all binomial $ps < .0001$), and this was true for each condition separately (Single and Multiple estimate; all binomial $ps < .01$). Thus, although the youngest children in our study were relatively inaccurate estimators, they nonetheless produced highly ordinal responses.
Figure 3.2: Proportion of ordinal responses for 5-, 6-, and 7-year-olds. Dashed line indicates performance in the Multiple Estimate condition; solid line indicates performance in the Single Estimate condition. Dotted line indicates chance performance. Error bars are SEM.

Our analysis of ordinality revealed that all age groups provided highly ordinal estimates. Importantly, this high level of ordinality did not simply reflect high levels of accuracy. While there was a strong relation between estimation error and ordinality in our dataset, only about half of the variability in ordinality was accounted for by estimation error (PAE; see Figure 3; $R^2 = .51$ for 5-year-olds; .68 for 6-year-olds; .38 for 7-year-olds). The same pattern – that typical measures of estimation accuracy only account for part of the variability in ordinality – was also found for other measures of estimation proficiency like estimation slope, linear $R^2$ (the $R^2$ of the linear fit between number being estimated and a participant’s estimate), and log $R^2$ (the $R^2$ of the logarithmic best-fit between number being estimated and a participant’s estimate). The accuracy measure that accounted for the most variability in ordinality scores was $R^2$ for the participant’s linear
fit ($R^2 = .70$ for 5-year-olds; .78 for 6-year-olds; .60 for 7-year-olds). Other accuracy measures were less predictive of ordinality (absolute value of the distance of linear slope from 1: $R^2 = .46$ for 5-year-olds; .73 for 6-year-olds; .32 for 7-year-olds; log $R^2$: $R^2 = .59$ for 5-year-olds; .69 for 6-year-olds; .68 for 7-year-olds). Thus, probably because ordinality does not require accuracy, it picks out variability that is not captured by existing measures of estimation ability.

The dissociation between estimation accuracy and ordinality is most evident when examining the performance of individual participants. For example, the 7 participants with the highest estimation error (5 5-year-olds and 2 6-year-olds) all had above-chance ordinality scores, and some participants with perfect ordinality scores nonetheless showed substantial estimation error (of the fifteen participants with perfect ordinality scores, participants’ average PAE ranged from just under 5 to just over 28, with a group mean of 12.37). Among the 44 participants who had very high levels of estimation error (PAE of between 15 and 30), 14 had ordinality scores greater than 95%, while only 4 of these participants showed ordinality performance below or near chance. Thus, while estimation error and ordinality were related, they are two fundamentally distinct phenomena (see Fig. 3.3).
Figure 3.3: Proportion of ordinal responses plotted by four different estimation measures. Each point is a child; triangles are 5-year-olds, circles are 6-year-olds, and X’s are 7-year-olds. Solid regression line indicates best fit for 5-year-olds, dashed line indicates best fit for 6-year-olds, and the patterned line indicates best fit for 7-year-olds. Dotted line at .5 is chance ordinality performance.

While ordinality was generally very high, most children made at least some estimates that were non-ordinal. We next asked whether these non-ordinal responses were randomly distributed across the number line (e.g., due to poor memory for previous responses or other random errors) or were instead restricted to particular numbers (e.g., due to relative unfamiliarity with large and less familiar numbers). To test this, we analyzed the relation between the numerical magnitude being requested on a particular trail and the likelihood that a participant provided an ordinal response. There was no effect of numerical Magnitude for any of our age groups (5 YOs: $\beta = -.005$, SE = .003, $p = .10$; 6 YOs: $\beta = -.004$, SE = .004, $p = .32$; 7 YOs: $\beta = -.003$, SE = .006, $p = .96$),
suggesting that children were no less likely to provide ordinal estimates for large
numbers than for small numbers (see Fig. 3.4).

Figure 3.4: Proportion of ordinal responses, plotted by the number being estimated. Each
point is the group mean for that number; triangles are 5-year-olds, circles are 6-year-olds,
and dots are 7-year-olds. Solid regression line indicates best fit for 5-year-olds, dashed
line indicates best fit for 6-year-olds, and the patterned line indicates best fit for 7-year-
olds. Dotted line at .5 is chance ordinality performance.

Together, these analyses reveal four properties of the development of reference
point selection and use during number-line estimation tasks. First, the fact that ordinality
was above chance for all age groups and in all conditions and independent of accuracy
suggests a reliance on previous trials to calibrate subsequent estimates. Second, these
analyses show that children make ordinal comparisons, even at the earliest ages tested,
and even in cases where their estimates were otherwise inaccurate. Third, older children’s
estimates were more likely to be ordinal than younger children’s estimates. Fourth,
ordinality was unrelated to the numerical magnitude of the number being estimated,
suggesting that limitations to children’s ordinal responses do not stem from lack of
knowledge of a particular numerical range, but instead from other factors (e.g., the accessibility of previous estimates, discussed below) that affected all trials equally.

While we have shown that children attend to the ordinality of numbers when estimating, one might wonder whether a child who relied only on trial-to-trial ordinality could nonetheless provide estimates that were linearly related to the target number being estimated (as our participants did). This might be of special importance for five-year-olds, whose rates of ordinal responding – while above chance – were low. Can participants who only provide ordinal estimates 69.9% and who do not rely on relative-distance information at all actually provide estimates that are linearly related to the target number being estimated? To address this concern, we turn now to our simulated estimation data. As described above, we ran a simulation of estimation data generated via trial-to-trial ordinality. This simulation assumed rates of ordinal responding that corresponded to those of our participants: For 5-year-olds, our simulation generated its 1st estimate completely randomly, and then provided ordinal estimates 69.9% of the time. We then asked whether these simulated 5-year-olds provided estimates that were linearly related to the target number being estimated – they did in 96.1% of simulations. Even in a model that took into account Distribution Condition (see analyses below for greater detail), 72.6% of our simulated experiments showed a linear relation between estimate and magnitude being estimated (Fig. 3.6). This suggests that trial-to-trial ordinality alone is sufficient to support the generation of estimates that are linearly related to the target magnitude, and that this is true even in populations with relatively low rates of ordinality. In fact, measures of estimation variability (computed by taking the mean for each estimate across all iterations of a simulation) appear relatively similar in our real 5-year-
old data and our ordinality-based simulated 5-year-old data (Simulated Linear $R^2 = .40$; Real Linear $R^2 = .57$), although even among 5-year-olds, the ordinality-only simulation does not appear to account for estimates of small numbers (under about 10). Subsequent simulations found that 6- and 7-year-olds’ ordinality rates led to a linear relation between target magnitude and estimate 99.8% and 100% of the time respectively, although the estimates were substantially more variable than the responses we found in our actual data (see Figures 3.5 and 3.6). These simulations show that it is possible for estimates generate via trial-to-trial ordinality alone to be linearly related to the target number requested, and – in some cases – to be relatively similar to children’s actual data.

Figure 3.5: Estimation performance for our Simulation of ordinality-based responses (top row) and Real Data (bottom row). Simulated data represents estimates generated by providing ordinal estimates at the rates typical of each age group; data in the top row represents the mean of all 1000 simulated experiments in a given age range; data in the bottom row represents the means for all real participants in a given age range. Triangles (left column) are 5-year-olds, circles (middle column) are 6-year-olds, and X’s / squares (right column) are 7-year-olds. Filled markers indicate Large Number Distribution data; unfilled markers indicate Small Number Distribution data.

Use of relative distance information
As noted earlier, to provide ordinal estimates, children only need to represent the relative ordering of the numbers. However, when adults make number-line estimates they express both ordinal and relative-distance relations between previous trials and later trials (Sullivan et al., 2011; Vul et al., 2013). And, as shown in Figure 5, while ordinality explains 5-year-olds’ data relatively well, simulations based on ordinality alone are more variable and less accurate than our real 6- and 7-year-olds’ data. This raises the possibility that an important developmental change in estimation ability is the shift from reliance on primarily ordinality-based analogical comparisons to relative-distance comparisons.

Accordingly, we assessed whether participants were sensitive to our Distribution manipulation. As described in the Introduction, previous research has shown that adults’ estimates are affected by similar Distribution manipulations: Adults who estimate from a Distribution that contains a disproportionate number of small numbers tend to make different estimates across the entire estimation task than those who estimate from a Distribution containing a disproportionate number of large numbers (Sullivan et al., 2011).

The basic logic behind the Distribution manipulation is that – if participants use their previous estimates to constrain subsequent estimates – then differences in estimates that occur early in an estimation task will carry through to later trials. For example, consider two participants: one is in the Large Number Distribution condition and is asked to estimate ‘96’ relatively early in the estimation task (and is statistically likely to underestimate it, perhaps as 90), and the other is in the Small Number Distribution asked to estimate ‘4’ (and is statistically likely to overestimate it, perhaps as 6). On the next
trial, both participants are asked to estimate ‘48’. If both participants rely on their previous estimates and do so by focusing on relative distance, then the participant in the Large Number Distribution should place their estimate of ‘48’ at 45 (since 96 : 48 :: 90 : 45), and the participant in the Small Number Distribution should place their estimate of ‘48’ at 72 (since 4 : 48 :: 6 : 72). In this way, small errors that are statistically likely to emerge based on the magnitude of the number being estimated (e.g., if you’re estimating ‘4’, there are more ways to accidentally overestimate it than to accidentally underestimate it) will carry through throughout an entire estimation task, but only if previous estimates constrain later ones, and do so via the tracking of trial-to-trial relative distance information.

Critically, the largest Distribution effects are found at the midpoint of the line. This is for two separate, but related reasons. First, estimates that are near the endpoints of the line are more constrained (e.g., I can underestimate ‘50’ by 10, but I can’t underestimate ‘7’ by 10). Thus, when making estimates near the endpoints of the number-line, it will not always be possible to carry error through from a previous trial, due to the constraints imposed by having fixed endpoints on the number line. As a result, we expect smaller effects of Distribution near the endpoints (see Sullivan et al., 2011 for evidence of this in adults). Second, the effect of Distribution is most likely if participants use their previous estimates to constrain later ones, and if they do so by representing relative distance information (see Fig. 3.6)\textsuperscript{15}. But, participants are less likely to use

\textsuperscript{15} It is possible – though substantially less likely -- for effects of Distribution to emerge if the participant uses previous estimates to calibrate later ones, but does so via ordinality alone. This is because, if the Large Number Distribution participant described above estimates ‘96’ as 90, then their estimate of ‘48’ cannot be larger than 90 (the expected
previous estimates to constrain estimates that are near experimenter-provided reference points (in this age range, the relevant reference points are the two endpoints; see Barth & Paladino, 2011; adults’ reference points are slightly different, see Sullivan et al., 2011). These two observations suggest that – insofar as children use their previous estimates to constrain later ones via relative-distance comparisons – we should find an Interaction of Distribution and Magnitude, such that the effect if Distribution is largest near the midpoint of the line.

We first conducted an analysis predicting children’s estimates from Magnitude, Age, Distribution, and all of their interactions. This analysis found an effect of Magnitude ($\beta = -.83, \text{SE} = .17, p < .0001$), an effect of Age ($\beta = -14.02, \text{SE} = 2.21, p < .0001$), an interaction of Magnitude and Age ($\beta = .23, \text{SE} = .05, p < .0001$), a marginal effect of Distribution ($\beta = -34.41, \text{SE} = 19.80, p = .08$), and no other effects. For this analysis, the absence of a significant interaction between Distribution and Magnitude does not indicate that there was no effect of Distribution: the effects of our Distribution manipulation were expected to be small (Sullivan et al., 2011) and may have been overwhelmed by the large differences in estimation performance we expect across age groups. Further, the Distribution effect is unlikely to appear if some age groups relied heavily on ordinality value, as calculated by averaging all possible estimates smaller than 90, is 45). In contrast, the participant in the Small Number Distribution who estimated ‘4’ as 6 would need to provide an estimate larger than 6 to preserve ordinality, and the expected value for this estimate would be 53.5. In this way, even participants who rely on ordinality alone could be influenced by Distribution: in our simulation based on 7-year-olds’ ordinality rates, an Interaction of Distribution and Magnitude emerged in 13% of experiments, while in our simulation of 5-year-olds’ data, only 5.9% showed an Interaction).
(especially if their ordinality scores were low). Because of this, we also conducted planned analyses separately by age group.

In these analyses, we predicted Estimate from Magnitude, Distribution, and their interaction. Predictably, all age groups showed an effect of Magnitude, demonstrating that their estimates were linearly related to the number they were estimating, even when Distribution condition was taken into account (5-year-olds: $\beta = .34$, $SE = .05$, $p < .0001$; 6-year-olds: $\beta = .50$, $SE = .03$, $p < .0001$; 7-year-olds: $\beta = .79$, $SE = .03$, $p < .0001$).

However, not all groups were affected by Distribution information. For 5-year-olds, there was no significant effect of Distribution ($\beta = -6.53$, $SE = 4.68$, $p > .15$) and no interaction between Distribution and Magnitude ($\beta = .002$, $SE = .06$, $p > .95$). Six-year-olds showed an effect of Distribution ($\beta = -16.25$, $SE = 4.98$, $p < .01$) and an interaction of Distribution and Magnitude ($\beta = .23$, $SE = .06$, $p < .0001$), suggesting that their estimates differed as a function of both the numerical Magnitude being estimated and the particular Distribution they were estimating. Seven-year-olds also showed an interaction of Distribution and Magnitude ($\beta = -.11$, $SE = .04$, $p < .025$), but no main effect of Distribution ($\beta = 3.99$, $SE = 4.80$, $p > .25$).

While it is clear that 6- and 7-year-olds showed an Interaction of Magnitude and Condition, one might wonder whether the best explanation of these data is really that older children use trial-to-trial relative distance information to calibrate their estimates. To further probe the source of the Interaction between Magnitude and Condition, we turn to our simulated data. When we simulated 1000 experiments in which participants used their previous estimate to calibrate later ones and did so by attending to relative distance, an interaction of Distribution * Magnitude emerged 69.1% of the time. No other
simulation yielded such high rates of Interactions. While it is theoretically possible to find an interaction in data generated via any of the estimation strategies that we simulated, based on the fact that we found significant interactions in both groups of older children, and based on the data provided by our simulation, the most likely explanation seems to be that children used their previous estimates to calibrate later ones, and did so based on the relative distance between estimates. This makes sense – we know that adults use previous estimates to calibrate later ones (Sullivan et al., 2011; Vul et al., 2013) and that children’s estimates of this numerical range by age 7 approach adult-like accuracy. One parsimonious explanation of these data is that older children use similar strategies to those used by adults when estimating.
Availability of Previous Estimates

We have shown that children make use of their previous estimates when estimating the location of a target number by making a comparison between previous estimates and the target number. In order to do this, they must be able to access the numerical magnitude and location of previous estimates. However, in typical estimation tasks, estimates are produced one-at-a-time, and the previous estimate is not visually available to the participant. This raises the possibility that children’s ability to produce accurate estimates may be limited by their ability to access information from previous estimates.
To test this, we constructed a model predicting participants’ estimates from the numerical Magnitude being estimated, Age, and Visual Access Condition (Single Estimate vs. Multiple Estimate). This model asks whether (a) individual estimates are predicted by the target number; (b) whether this relation is modulated by having visual access to previous estimates and (c) whether this effect was mediated by age. Critically, an interaction of Condition and Magnitude (meaning that participants gave smaller estimates for smaller numbers and larger estimates for larger numbers in one of the two conditions) would provide evidence that having visual access to previous estimates improves estimation performance. Predictably, we found an effect of Magnitude ($\beta = -.43$, SE = .14, $p < .01$), an effect of Age ($\beta = -9.23$, SE = 2.16, $p < .0001$), and an interaction of Age and Magnitude ($\beta = .17$, SE = .02, $p < .0001$). Again, this shows that participants provided estimates that were related to the target magnitude, and that older children made more accurate estimates than younger children. Critical to our main question, we also found an interaction of Condition and Magnitude ($\beta = -.56$, SE = .22, $p < .05$), suggesting that participants in the visual access condition provided more accurate estimates than participants in the no access condition. We also found a three-way interaction of Condition, Age, and Magnitude ($\beta = .09$, SE = .04, $p < .05$), suggesting that the benefit of having visual access to previous estimates was restricted to children of a certain age.

To better understand these effects, we next analyzed each age group separately. Again, across all age ranges, Magnitude remained a significant predictor of estimation performance (5-year-olds: $\beta = .41$, SE = .04, $p < .0001$; 6-year-olds: $\beta = .55$, SE = .04, $p < .0001$; 7-year-olds: $\beta = .75$, SE = .03, $p < .0001$). While older children performed better
than young children, at each age group participants’ estimates were related to the target number being estimated.

Critically, only five-year-olds showed an interaction of Condition and Magnitude ($\beta = -.18, \ SE = .06, p < .01$): these youngest participants were more likely to provide smaller estimates for smaller numbers and larger estimates for larger numbers in the Multiple Estimate condition (where it was easy to access previous estimates) relative to the Single Estimate condition (where it was hard to access previous estimates). This resulted in more accurate performance in the Multiple Estimate condition than in the Single Estimate condition (slope relating estimates to the target number for Single Estimate condition: $\beta = .21$; Multiple Estimate condition: $\beta = .38$; here, $\beta$ can be interpreted as a simple slope measure with adult-like performance as $\beta = 1$). We found a similar benefit in 5-year-olds for other measures of estimation accuracy (PAE: Single Estimate = 28.4 and Multiple = 23.58; linear r-squared: Single Estimate = .16 and Multiple = .34 log r-squared: Single Estimate = .19 and Multiple = .39).

Six-year-olds showed a main effect of Condition ($\beta = -8.50, \ SE = 4.30, p < .05$), but no interaction ($\beta = .05, \ SE = .05, p > .25$). This suggests that participants in the Multiple Estimate condition made, on average, larger estimates than those in the Single Estimate condition, but without an interaction, we cannot conclude that having visual access to previous estimates improved 6-year-olds’ estimates. Finally, 7-year-olds showed neither an effect of Condition ($\beta = -.48, \ SE = 4.81, p > .9$), nor an interaction ($\beta = -.008, \ SE = .04, p > .8$). This suggests that these older children, whose estimates tended to be quite accurate and linear even in a standard number-line estimation task (e.g., Booth &
Siegler, 2006), do not show improved performance from having visual access to previous estimates, likely because their performance was already quite strong.

Because our 5-year-olds were the least likely to provide accurate estimates, but still seemed to rely heavily on previous estimates when estimating (as evidenced by the effect of Visual Access condition and by their above-chance rates of ordinal responding), we asked one final question of our data. We asked whether these children not only provided estimates that were in the correct direction (ordinal) relative to the immediately preceding estimate \((n-1)\), but also relative to all previous estimates \((n-2, n-3 \ldots n-11)\). As in our initial ordinality analysis, we coded each trial as 1 if it was in the correct direction relative to the previous trial \(n-x\), and as 0 if it was in the incorrect direction. We did this independently for each of the 11 preceding trials (so that a failure to provide an ordinal response on trial \(n-8\) relative to trial \(n\) did not preclude a successful ordinal response on trial \(n-9\) relative to trial \(n\)). Importantly, 5-year-olds’ estimates were more ordinal relative to each of the 11 previous trials tested than would be expected by chance (50%) alone (all binomial \(p s < .001\)). For each of the 11 previous trials, responses in the Multiple Estimate condition were more likely to be ordinal than those in the Single Estimation condition (see Fig. 3.7), although this pattern did not reach significance for all comparisons, likely because power decreased as the distance from \(n\) increased.\(^{16}\) Together, the effects of Condition on accuracy and ordinality suggest that even our youngest participants used

---

\(^{16}\) Chi-squared tests found significant effects of Single vs. Multiple estimates on ordinality at: \(p = .093\) for \(n-2\), \(p = .025\) for \(n-3\), \(p = .058\) for \(n-4\), \(p = .032\) for \(n-5\), \(p = .092\) for \(n-6\), \(p = .26\) for \(n-7\), \(p = .29\) for \(n-8\). However, computations of ordinality are subject to increasing data-loss as the trial distances become larger (e.g., to compute ordinality for \(n-11\), comparisons can only be made for data points after trial 11, whereas \(n-1\) comparisons can make use of all data except for the 1st trial). Thus, non-significant tests involving larger intertrial distances may be underpowered.
their knowledge of ordinality to relate the target number being estimated to all previous trials, and that they are best able to do this when they have visual access to their previous estimate.

![Graph](image)

**Figure 3.7:** Proportion of 5 year-olds’ ordinal responses plotted by trial, where \( n-1 \) represents the proportion of responses on trial \( n \) that are ordinal relative to trial \( n-1 \), and where \( n-7 \) represents the proportion of responses on trial \( n \) that are ordinal relative to trial \( n-7 \). Filled triangles are data from the Single Estimate condition; carats are from the Multiple Estimate condition.

**Discussion**

In this study, we investigated whether the analogical reasoning demands of estimation tasks influence children’s number-line estimation performance. We reasoned that, if number-line estimation is fundamentally analogical, then estimation should require selecting a numerical reference point (e.g., a point on the line or a previous
estimate), identifying a relation between that reference point and the target number being estimated, and extending this numerical relation to the spatial domain. To how the analogical demands of number-line estimation tasks affect the developmental trajectory of estimation, we asked (1) which reference points children use and (2) how they use these reference points when estimating. We showed that by age 5, although children are very inaccurate estimators, they are also not random, and have begun to use analogical processes to guide their estimates. Specifically, we found that children attend to the numerical ordering of the target number and their previous estimates – and use this ordering information when estimating. We also found that number-line estimation changes (1) as children learn to identify and order reference points on the number line, (2) as they learn to represent the relative distance between these points, and (3) as they become better able to access previous estimates when they are not otherwise visually available. These data suggest that important changes in the development of number-line estimation are related to the analogical demands of number-line tasks.

Counter to the view that poor estimation performance reflects a lack of familiarity with a particular numerical range (e.g., Lipton & Spelke, 2005; Siegler & Opfer, 2003), we found that even our youngest participants made ordinal responses – meaning that their estimates were in the correct direction relative to previous estimates – and that they did so even when they failed to make accurate estimates. Amazingly, this was true at all magnitudes tested, suggesting that children made use of ordinality across the entire number-line. This finding demonstrates that very young children may have a greater understanding of the relation between numbers within the count list than had previously been documented (see also Barth et al., 2009; Sullivan & Barner, in press).
We also showed that the accuracy and ordinality of 5-year-olds’ estimates improved when they had visual access to previous estimates. This suggests 5-year-olds’ difficulties with estimation are not purely due to a lack of numerical knowledge, but are also partially the result of difficulty accessing previous estimates. As noted earlier, a critical component of the analogical comparison required by estimation is relating a target number to a reference point. If children use their previous estimates as reference points, then their ability to access these previous estimates could limit estimation performance. Consistent with this, we found that our youngest participants made better estimates when they had visual access to their previous estimates. Critically, having visual access to previous estimates should not have influenced estimation performance if children didn’t use their previous estimates as points of comparison. While the present study did not directly test which cognitive mechanisms limited our youngest participants from using their previous estimates when these estimates were not visually accessible, possible explanations include working memory, domain-general processing ability, and the ability to update mental representations. Future studies should explore these possibilities.

We also found that attested changes to estimation performance between the ages of 5 and 7 (e.g., Ziegler & Opfer, 2003; Booth & Siegler, 2006) may stem in part from developmental changes to children’s ability to construct sophisticated analogical numerical comparisons. For example, we found that the ability to make comparisons on the basis of the relative distance between numbers (and not just their ordinality) develops gradually between the ages of 5 and 7. Evidence for this came from two sources. First, our simulated estimation data showed that estimates generated by following trial-to-trial ordinality alone are relatively similar to 5-year-olds’ actual estimates, but are relatively
dissimilar to the estimates provided by our 6- and 7-year-olds. Second, we found evidence that the distribution of numbers that children estimated from influenced 6- and 7-year-old’s estimates, but not 5-year-olds’ estimates. The most likely explanation of this finding is that only the older children calculated the relative distance between previous estimates and new estimates. Taken together, these data suggest that improvements in estimation performance between the ages of 5 and 7 stem from multiple sources. Older children not only are better at accessing and manipulating previous estimates, but they also attend more to the relative distance between numbers (and not just ordinality).

While this study showed that children make number-line estimates by relating a target number to numerical reference points (e.g., previous estimates), several questions remain about the development of this process. First, the cause of the shift from using information about ordinality to making estimates on the basis of relative distance information remains unknown. On the one hand, children might have knowledge of the relative distance between numbers from a very young age, but simply be unable to deploy this knowledge. This might be because they fail to appreciate its relevance to the task, because respecting distance relations imposes addition domain general demands on working memory, attention, etc. (e.g., Anobile et al., 2013), or because children initially have trouble transferring information regarding the distance between numerals to the spatial domain. On the other hand, children may initially lack sophisticated knowledge of the relative distance between numbers, and only acquire this knowledge after prolonged exposure to the number system and counting routine. Consistent with this, it seemed that even our 5-year-olds used relative distance information for the smallest numbers estimated (see Fig. 3.4). These two possibilities point to very different mechanisms of
estimation change, and to different pictures of early numerical knowledge. Future research should disambiguate between these two possibilities.

Another question left open by our study – and our result that 5-year-olds’ estimation performance is affected by access to past estimates – is how memory capacity and number-line estimation performance are related. Recent studies, like ours, suggest that estimation and memory for numbers may be related, but leave open the mechanism that governs this relation. In one study, Thompson and Siegler (2010) showed that children who had better memory for number words were more likely to be linear estimators – a result which was taken as evidence that memory for numbers is improved by the development of an adult-like system of number representations. However, another explanation for these data is that children who have better memory for numbers may be better able to provide accurate estimates, because estimation involves analogical processes which depend critically on memory – i.e., accessing and manipulating past estimates. While our data do not directly test the role of working memory in estimation, our finding that 5-year-olds performed better when they had visual access to their previous estimates is consistent with the idea that memory for numbers is a prerequisite for estimation rather than a product of mature number representations (see Bryant & Trabasso, 1971; see also Siegler & Booth, 2004).

This study suggests that structural analogy is likely fundamental to number-line estimation: Children compute within-domain numerical relations between target numbers and numerical reference points when estimating, and then analogically extend these relations to the spatial domain (marking their numerical estimates on a spatial number line). While past reports have detailed the cognitive consequences of the spatial
comparisons required by this analogy (Cohen & Blanc-Goldhammer, 2011; Barth & Paladino, 2011; Slusser et al., 2013), ours is the first to test the numerical component of this analogy, and to show that children use their own previous estimates as reference points. We showed that children as young as 5 make analogies when estimating, and that both (1) the types of relations that guide these analogies and (2) the ease with which children compute these analogies change over development.

These findings not only contribute to an explanation of the developmental transitions in estimation ability described in this study, but are also consistent with past findings in the literature. For example, our data are broadly consistent with reports that young children use experimenter-provided reference points (e.g., midpoint and endpoint) to calibrate their estimates, and that children’s reference point use affects their estimation (Barth & Paladino, 2011; Slusser et al., 2013; Sullivan et al., 2011). Also, while these data suggest that estimation tasks depend on non-numerical capacities required by analogy, they are also consistent with a role for developing numerical capacities which change in parallel (e.g., Siegler & Opfer, 2003). Thus, our study suggests that estimation is a multi-faceted process that draws on multiple cognitive capacities (Thompson & Opfer, 2010), and which therefore cannot be interpreted as a simple, pure, proxy for the development of numerical representations.

In summary, we have argued that number-line estimation tasks are fundamentally analogical: Estimation depends critically on the ability to relate a target number to a previous estimate and then extend this relation to the spatial domain. Our data show that, due to the analogical component of number-line estimation tasks, estimation abilities may be limited by factors that are not directly related to numerical knowledge. This suggests
that previous work may have underestimated young children’s estimation abilities, and that 5-year-olds may possess more sophisticated knowledge of the number system than previously believed. In addition, these findings paint a new picture of the mechanisms underlying estimation performance by demonstrating that analogical comparison guides young children’s estimates.
Chapter 3, in full, is a reprint of the material currently under review at the *Journal of Experimental Child Psychology*. Sullivan, J., & Barner, D. The dissertation author was the primary investigator and author of this paper.
References


GENERAL DISCUSSION

Humans have two systems for representing number: a nonverbal approximate system (ANS) and a verbal number system. While the ANS is available to infants from birth, the verbal number system is acquired over the first few years of life. By age 5 – and perhaps before – children have begun to connect their verbal and nonverbal number systems together. However, as I have shown, the mappings between these two systems continue to be refined during the elementary school years. This raises a question: how do the verbal and nonverbal number systems become related to one another in development? This dissertation tested the learning mechanisms that children recruit to connect the verbal and nonverbal system to one another.

In Chapter 1, we investigated the roles of two learning mechanisms in guiding the mappings between the verbal and nonverbal number systems: Associative Mapping (AM) and Structure Mapping (SM). AM is an item-by-item association of a particular word (e.g., fifty) onto a particular numerical magnitude (e.g., ‘about 50 things’) via experience. SM, in contrast, is a holistic inference about the relation of the verbal and nonverbal number system, made on the basis of their similar structures.

We found converging evidence that adults likely rely on AM for mappings up to about 12. One source of evidence for this conclusion came from adults’ performance on our Number Matching task, where we found that participants were highly accurate at matching a number word to one of two visually presented arrays, provided that the smaller of the two arrays contained 12 or fewer items. A second source of evidence came from the Calibrated Estimation task, in which we found that adults’ estimates of small
arrays were resilient to misleading feedback, suggesting that AM supported mappings for arrays of that size.

We also found evidence that participants relied on SM for larger number word mappings. The strongest evidence for this conclusion was that after hearing misleading feedback about an individual mapping (e.g., “the largest number you’re going to see is 750” when it was really 350), participants’ estimates were altered for all other numbers, suggesting that number word mappings are mutually constraining.

While Chapter 1 tested adults’ number word mappings, Chapters 2 focused on understanding the formation of these mappings, and how they change during early childhood. In order to understand the development of children’s number word mappings, we tested 5- to 7-year-olds on the same tasks described in Chapter 1, in order to assess their reliance on AM and SM. In addition, we measured children’s verbal number knowledge with a counting assessment and test of their knowledge of the relative ordering of the number words.

At the outset, we had several hypotheses about the ways in which AM and SM might interact during development. One possibility is that children initially rely on AM to form number word mappings, and don’t recruit SM at all until they are older. This is a plausible developmental trajectory because children could form individual mappings via AM on the basis of a few item-specific experiences and in the absence of knowledge about the structure of the count list as a whole. An alternative possibility is that children must amass a sufficiently large set of AMs prior to deploying SM. By this view, children might form AMs up to about 12 (like adults do), and then begin forming mappings via
SM. A final possibility is that SM is the primary mechanism by which children connect the verbal and nonverbal number systems, and that adult-like AMs form due to experiences with particular word-magnitude pairings – not because AM is the primary mechanism responsible for forming number word mappings.

We found that both 5-year-olds and 7-year-olds recruited SM for all mappings above about six, and that our measures of verbal number knowledge did not predict children’s reliance on AM or SM. However, we also found that participants in our Calibrated Estimation task showed a much larger (and more adult-like) effect of calibration at age 7 than at age 5. Taken together, these data suggest that children have sufficient knowledge of the number system by age 5 to form mappings via SM, and therefore that structurally-based inferential processes are important for supporting word-world mappings. In addition, the fact that calibration effects differed across ages suggests that there may be important developmental refinements to the precision of mappings supported by SM, an idea that we explored further in Chapter 3.

In Chapter 3, we studied children’s use of SM more explicitly. In particular, we developed a model of estimation that took as its premise that number-line estimation tasks are fundamentally analogical reasoning tasks. Doing this allowed us to isolate the steps required to make structural inferences about the relation between number language and nonverbal representations of number. We could then ask (1) whether children construct number word mapping via trial-to-trial inferences during typical estimation; (2) whether children use their previous estimates to calibrate later estimates during typical estimation; (3) whether children attend to the ordinal structure of the count list or to the
structure of the relative distances between numbers; and (4) whether the analogical reasoning demands of typical estimation tasks may limit young children’s ability to excel at estimation.

We found that children between the ages of 5 and 7 used their previous estimates to calibrate later ones. Importantly, we were also able to better understand the type of structural inference that children made about the number system when forming their mappings. We found that 5-year-olds attended more to the ordinal structure of the count list, while 6- and 7-year-olds attend to the relative distance between numbers. In addition, we found that the task demands imposed by needing to remember previous estimates in order to calibrate later estimates harmed 5-year-olds’ estimation ability. These data suggest that estimation tasks do not simply test number knowledge – they test a host of non-numerical skills. In addition, these data suggest that developmental changes in estimation ability stem in part from refinements to the types of structures that children pick out when forming mappings via SM.

While we showed in Chapter 3 that estimation tasks are fundamentally analogical and require structural inference, it is still unknown how these findings relate to the development of children’s domain-general analogical reasoning skills. While number-specific structural knowledge is required to support mature SM (e.g., that ‘40’ is exactly twice as large as ‘20’), one possibility is that the refinements to SM that we see during the early elementary-school years are not restricted to the domain of number – instead, children become better during these years at analogy – in general. Of particular interest is the ability to entertain multiple different possible analogical relations within a given
domain (e.g., ordinality; relative distance). In this dissertation, I have shown that adults and older children are able to pick out multiple structural relations between number words in the count list, and then these relations to form mappings via SM. Is the development of the ability to pick out multiple numerical relations within the count list best explained by domain-specific changes to number knowledge, or to domain general changes to analogical reasoning skill? One reason that we might want to know whether estimation performance improves because of (a) improvements to domain-general analogical skills or (b) improvements to domain-specific knowledge of the structure of the count list is that estimation ability is related to academic success. Thus, a critical next step is to understand the relation between estimation skill, math success, and domain-general analogical reasoning skill.

A second question that remains is whether the inferential processes described within this thesis support the connection of language to nonlinguistic representations outside of the domain of number. The verbal number system is highly structured, making it relatively easy to define a structure within the verbal number domain (e.g., relative order; relative distance) and then use that structure to make a holistic inference about the relation between the verbal and nonverbal number domains. However, it seems entirely possible that the vast majority of words that children learn are not situated within systems that have such highly transparent lexicalized structures. In fact, besides classes of words like colors, familial relations, some taxonomies, time words, and quantifiers, relatively few words belong to structured, lexicalized groups. How do we map words onto our experiences with the world in the absences of these lexicalized structures? One of my
ongoing lines of research begins addressing this question by asking how children use discourse structures to isolate the referents of words.

A third question that remains is the extent to which the inferential processes described in this dissertation shape children’s use of number words in daily life. It is possible to use number words not to denote precise (or even approximate) quantities, but rather to place bounds on possible quantities under discussion. For example, the interpretation assigned to twenty differs in the sentences: “I counted them and I have exactly twenty dollars” [exact]; “There were twenty people at the party” [approximate]; “You can use the bus pass twenty times per month” [upper-bounded]; and “People with A+’s in twenty classes will be on honor roll” [lower-bounded]. While this dissertation – by positing that inferential processes are fundamental to the process of giving content to number words – is consistent with the types of mechanisms likely required to use and interpret number language in the examples above, the version of SM described in this dissertation is not sufficient to explain the range of content that can be assigned to number words. Future research will be required to understand the nature of the inferences that children and adults draw when interpreting number words in naturalistic speech.

A final question that remains is about how mappings between number language and nonverbal representations of number are organized in people who have access to multiple languages (e.g., bilinguals). This question becomes especially important when considering languages that differ in how they structure the count list. For example, the literal translations of the English numbers [sixty, seventy, eighty, and ninety] are [sixty,
sixty-and-ten, four-twenties, and four-twenties-and-ten] in some dialects of French, and
[six-ten, seven-ten, eight-ten, and nine-ten] in Mandarin. If we map number words onto
numerical quantities via holistic inferences based in the structure of the count list we use,
then how do cross-linguistic differences in the structure of count lists influence
estimation ability?

While there are several outstanding – and potentially productive – research
questions left open by the work presented within this dissertation, when taken together,
this dissertation provides a crucial first step towards understanding the learning
mechanisms that children and adults recruit when connecting number language to
nonverbal representations of number.