UNIVERSITY OF CALIFORNIA SAN DIEGO

How language-specific experiences contribute to number concepts development; Evidence from multilingual learners

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

in

Experimental Psychology

by

Elisabeth Marchand

Committee in charge:

Professor David Barner, Chair
Professor Adena Schachner Brady
Professor Federico Rossano
Professor Caren Walker
Professor Eva Wittenberg

2022

The Dissertation of Elisabeth Marchand is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I owe the most profound gratitude to my advisor and mentor David Barner for his invaluable guidance, support, and contribution during the creation of this work. Throughout the years, David has set high expectations that facilitated my growth as a researcher and as a thinker, and I am infinitely grateful for his commitment to my scientific career.

I would also like to thank members of my committee, Adena Schachner Brady, Caren Walker, Eva Wittenberg and Federico Rossano for their helpful feedback, flexibility and encouragement. Their expertise and insights undoubtedly improved the quality of this work.

Thanks are due to members of the Language and Development Lab for their excellent feedback and insightful discussions. I am deeply grateful to have found through this lab a community with such inspiring and creative people.

I am also indebted to all the amazing research assistants and lab managers (Junyi Chu, Ashlie Pankonin and Kelly Kendro) for their hard work and contribution to each chapter of this thesis. This work would also not have been possible without the hundreds of families and children who kindly agreed to participate in my research.

I am very thankful to my co-authors on the paper that became Chapter 1, Shirlene Wade and Jessica Sullivan. Their expertise helped me develop my skills as a researcher when I first joined the lab. To my other co-authors Jarrett Lovelett and Kelly Kendro, who contributed to Chapter 2; learning as much in statistical reliability would never have been as enjoyable without the company of such good friends. To Nina Schoener, Jocelyn Sandoval-Franquez, and Kelly Kendro, my co-authors in Chapter 3, thank you for giving me the great pleasure of being your mentor and I am so thankful for your dedication to this work.

Thank you to Boutheina Jemel and Julie McIntyre, at the University of Montreal, who introduced me to research and encourage me to dream beyond what was attainable.

I am also indebted to my family in San Diego: Maddie, Jarrett, Drew, Tiffany, Jae, Brendan, Brian, Rocio, Meredith and Ross, and to friends in the UCSD Psychology graduate program. All these years would not have been so enjoyable without you.

I would also like to thank all my friends in Montreal and the Sainburgs for all their encouragement, patience, and support during this journey.

To my parents and brother, I would like to dedicate this thesis. You were always my most fiercefull advocates. Merci pour tout.

Finally, to Tim Sainburg, my best friend, husband and favorite Bird Nerd. He deserves a special recognition for his daily support and encouragements. No challenge is impossible to tackle with someone as kind, generous and loving to accompany you during the process.

Chapter 1, in full, is a reprint of the material as it appears in Marchand, E., Wade, S., Sullivan, J., and Barner D. (2020). Language-specific numerical estimation in bilingual children. *Journal of Experimental Child Psychology*, doi.org/10.1016/j.jecp.2020.104860. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Marchand, E., Lovelett, J. T., Kendro, K., and Barner, D. (2022). Assessing the knower-level framework: How reliable is the Give-a-Number task?. *Cognition*, 222, 104998. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Marchand, E., Schoener, N., Sandoval-Franquez, J., Kendro, K., and Barner, D. The dissertation author was the primary investigator and author of this paper.

VITA

2013 Bachelor of Science in Speech Pathology, University of Montreal

2014 Master Degree in Speech Pathology, University of Montreal

2015 Certificate in Psychological Sciences, University Catholic of Louvain

2022 Doctor of Philosophy in Experimental Psychology, University of California San Diego

PUBLICATIONS

Marchand, E., Schoener, N., Sandoval-Franquez, J., Kendro, K., & Barner, D. (in prep). The Development of Subitizing in Bilingual Children.

Marchand, E., Lovelett, J. T., Kendro, K., & Barner, D. (2022). Assessing the knower-level framework: How reliable is the Give-a-Number task? *Cognition*, 222, 104998.

Marchand, E., Wade, S., Sullivan, J., & Barner, D. (2020). Language-specific numerical estimation in bilingual children. *Journal of Experimental Child Psychology*, *197*, 104860.

Marchand, E., & Barner, D. (2019). The Acquisition of French Un. *Proceedings of the 41th Annual Meeting of the Cognitive Society*.

Marchand, E., Barner, D. (2018) Analogical Mapping in Numerical Development. In D. Berch, D.G., editor(s), *Language and Culture in Mathematical Cognition*, pages 3-15. New York: Elsevier.

Barner, D., Athanasopoulou, A., Chu, J., Lewis, M., Marchand, E., Schneider, R., & Frank, M. (2018). A one-year classroom-randomized trial of mental abacus instruction for first-and second-grade students. *Journal of Numerical Cognition*, *3*(3), 540-558.

ABSTRACT OF THE DISSERTATION

How language-specific experiences contribute to number concepts development; Evidence from multilingual learners

by

Elisabeth Marchand

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2022

Professor David Barner, Chair

Unlike other animal species, humans have the ability to represent large exact quantities. While different theories in number cognition have attributed this ability to our access to natural language, the question of how exactly natural language affords humans this unique ability remains unclear. Studying bilinguals provides a valuable approach to investigating the relationship between numbers and language, as documenting the similarities and differences across languages can inform us about the role of language-specific experiences in the development of numerical

representations. In this dissertation, I will argue that language-specific experiences play a fundamental role in the earliest steps of number acquisition, drawing on the evidence from 3- to 7-year-old bilingual children. In Chapter 1, I present evidence that French-English bilinguals estimate the numerosity of arrays of dots differently across their two languages. This asymmetry in bilinguals' mappings between number words and non-verbal representations across their two languages reveals that these mappings rely on language-specific knowledge of the structure of their count lists. In Chapter 2, I investigate some potential methodological issues when testing bilingual children and show that bilingual studies should take into account how test-retest reliability can contribute to observed differences across languages in bilinguals. However, in the case of Give-a-Number, some levels are more affected than others. Finally, in Chapter 3, I explore further the role of language-specific experiences in the mappings between number words and non-verbal representations by showing that children who know how to count do not subitize similarly across their two languages. Instead, differences in subitizing skills across languages suggest that language-specific experiences play a role from the very beginning of number word acquisition. Together, these studies suggest that language-specific experiences play a major role in building mappings between number words and non-verbal representations, via the estimation of large and small sets. These studies also reveal that some basic numerical abilities don't transfer across languages in bilinguals.

INTRODUCTION

Many animal species possess the ability to approximately perceive the numerical properties of sets in the world. However, only humans have created external symbolic representations of numbers like body counting systems, number words, written numerals, and the abacus, among others, that allow us to go beyond our noisy perceptions of magnitudes and capture exact numerical information about our environment (Ifrah, 2000; Menninger, 1969). These symbolic systems have provided us with the foundation to elaborate sophisticated concepts such as variables, matrices, and infinite numbers, which today help us understand the world we inhabit. Hence, the creation and transmission of early symbolic numerical systems represent an important tour de force of the imagination that changed the way we navigate our environment. What is the source of these achievements? Some researchers have suggested that having access to natural language plays an important role in affording us those unique abilities (Carey & Barner, 2019; Gordon, 2004; Le Corre & Carey, 2007; Pica et al., 2004; Spaepen et al., 2013; Spelke, 2017). However, the question of how exactly language comes to drive changes to our numerical representations remains unclear. Recently, some investigators have taken the approach of studying bilingual learners to explore the role of language in the development of number concepts (Spelke & Tsivkin, 2001; Wagner et al., 2015). This approach has the advantage of isolating the role of language from nonlinguistic factors that also drive changes to our numerical representation (e.g., executive function or maturation of nonlinguistic numerical representations), but that are shared across languages in bilinguals. Studying bilinguals, therefore, allows us to distinguish between language-specific knowledge and knowledge that is not specific to a particular language and can transfer across languages. In this thesis, I adopted this approach to investigate the role of language in the development of a basic

numerical ability, numerical estimation, that provides a unique window into how we come to associate our linguistic and nonlinguistic representations of number.

Nonlinguistic systems of magnitude representation

One clue substantiating the role of language in numerical cognition comes from studies of non-linguistic number systems. This work reveals that, in absence of symbols like number words, pre-verbal humans and non-human animals (e.g., fish, birds and rats) are unable to represent large exact numbers but can still apprehend magnitudes – in limited ways – using two nonlinguistic systems: the Approximate Number System (ANS; or Analog Magnitude System) and the Object Tracking System (OTS) (Dehaene, 2011; Feigenson, Carey, & Hauser, 2002; Feigenson, Dehaene, & Spelke, 2004).

The ANS is an evolutionarily ancient system that enables us to discriminate and compare the approximate magnitudes of sets (Barth, Kanwisher, & Spelke, 2003; Dehaene, 2001; Dehaene & Changeux, 1993; Meck & Church, 1983; Whalen, Gallistel, & Gelman, 1999). Previous studies of this system indicate that it encodes the magnitudes of sets perceived in the world by automatically creating analog mental symbols proportional to the number of individuals contained in the represented sets. In particular, representations in the ANS are governed by Weber's law, according to which the threshold of discrimination between two sets is a function of their numerical ratio. For example, it is easier to tell the difference between 5 vs 10 dots (1:2 ratio) than 40 vs 45 dots (8:9 ratio), even though in both cases, there is a difference of 5 dots (Lipton & Spelke, 2004; Piazza, 2011; Wilkey & Ansari, 2020; Xu, 2003). Previous studies in developmental psychology have shown that the acuity of the ANS (Weber's fraction) increases with age, especially within the first year of life. Specifically, 6-month-old infants can discriminate sets that stand in a 1:2 ratio

but not 2:3, whereas 10-month-old infants can discriminate both ratios, and adults can discriminate ratios of 7:8 or closer (Halberda & Feigenson, 2008; Lipton & Spelke, 2003; Xu & Spelke, 2000).

The second nonlinguistic system that has been argued to play a role in the perception of magnitude is the Object Tracking System (OTS; Feigenson & Carey, 2003, 2005; Feigenson, Carey & Hauser, 2002). This system allows us to keep track of small numbers of individual objects, in parallel, in working memory by creating mental symbols for each individual object in the sets. In one assessment of this system, 10- and 12-month-old children watched an experimenter hide crackers into two opaque buckets. The buckets contained different numbers of crackers and children spontaneously chose the bucket containing the largest number of crackers when the ratio of crackers was 1 vs 2 and 2 vs 3. However, children were at random when they had to choose the bucket with the largest number of crackers when the buckets contained 1 vs 4, 2 vs 4 or 3 vs 4 crackers (Feigenson, Carey & Hauser, 2002). Similarly, in manual search paradigm in which 14-month-old infants saw objects hidden sequentially in an opaque box, researchers observed that children's pattern of search matched the number of objects hidden but only when the box contained 3 objects or less. When 4 objects were hidden and 1 was retrieved, infants did not search for more. Together, these studies suggest that this system has a representational capacity limit of about 3 objects (Feigenson & Carey, 2003). However, similar to the ANS, there are individual differences in the acuity of this system and evidence of maturation during development (Piazza, 2011).

The literature reviewed above reveals that humans are endowed with two nonlinguistic systems that allow us to apprehend quantities in the world: the ANS and the OTS. However, both systems have important limitations; the ANS representations are approximate by nature and the OTS is limited to representing only a few individuals. In contrast to these nonlinguistic systems, humans have created linguistic number systems like the count list that allows us to represent exact

concepts such as “307”. During development, the nonlinguistic and linguistic systems of numerical representations become related and from this association, humans can engage in activities such as numerical estimation. In the current thesis, I explore the role that language plays in the development of numerical estimation. Studying the development of estimation is important for two reasons: first, past studies have shown that estimation is not only correlated with basic numerical skills such as arithmetic (Booth & Siegler, 2008; Siegler & Ramini, 2008, 2009) but it is also correlated with overall academic achievement in school-age children (Duncan et al., 2008; Jordan et al., 2009). Second, estimation is important because it is a process that is commonly used in everyday life. For example, we often have to estimate distances to travel from place to place and we try our best to estimate the time tasks will take us to perform. Estimation is also often used in research when precise enumeration is impossible. For example, in biology, when researchers are unable to count the exact number of individuals in a population, such as a murmuration of starlings, they rely on estimation to keep track of population growth or decline, which in turn can inform us of the impact of our actions on the environment. In this dissertation, I investigate specifically the role of language in the development of estimation abilities in bilingual children as a case study to help elucidate how language contributes to the development of basic numerical concepts.

Numerical Estimation

The process of estimation is the ability to attribute a numerical symbol, such as a number word, to a nonverbal representation of magnitude, such as a set of dots, rapidly and without counting. Typically, in an estimation task, participants are presented with flashed arrays of dots and are asked to provide a verbal estimate of the numerosity of these sets. The rapid presentation of dots is meant to prevent subjects from counting so that they have to rely on their intuitive sense

of how number words are represented by nonverbal magnitudes (and vice-versa). Previous studies of estimation of large numbers have shown that estimation accuracy decreases as the numerosity of arrays increases, unlike the estimation of small sets (sets of 1 to 4 dots), for which the accuracy remains constant across numerosity (Burr, Turi, & Anobile, 2010; Indow & Ida, 1977; Revkin et al., 2008). Because of this qualitative difference, the estimation of small sets is more commonly referred to as subitizing.

There are three components that underlie the process of estimation: First, subjects need to perceive the set. Second, subjects automatically create a mental representation of the magnitude of the set in the nonlinguistic systems (either ANS or OTS). Third, subjects need to retrieve and attribute a verbal label (i.e., a number word from the count list) to the set based on a mapping mechanism between the nonlinguistic and linguistic systems of numerical representations. Previous studies have shown that there are two types of mapping mechanisms involved in estimation: item-based associative mappings and structure mappings (Sullivan & Barner, 2013, 2014).

Associative mapping is a type of mapping in which a number word is connected directly on an item-by-item basis to the mental representation of a particular magnitude (Sullivan & Barner, 2012, 2014). For example, to be able to rapidly and accurately subitize a set of three dots, a participant would need to have the number word *three* associated to a long-term mental representation of *threeness.* Previous studies of estimation suggest that adult participants create these types of mappings mainly for small magnitudes (e.g., 1 to 8 dots; Sullivan & Barner, 2013). However, this type of mapping is important to the development of numerical concepts because dominant accounts of conceptual development have argued that it is by forming associative mappings that children learn the meanings of the first number words (Carey, 2004).

A large number of studies have now shown that children start making associative mappings between number words and cardinalities (numbers of elements in a set) around the age of 2, and when they do so, they learn these mappings in a protracted sequence (Sarnecka & Carey, 2008; Sarnecka & Lee, 2009; Wynn, 1990, 1992). Most of the findings in early number word acquisition come from studies that have used a task called Give-a-Number (Give-N), in which children are asked by an experimenter to give specific quantities of objects. Using this task, studies have shown that before children create any associative mappings, they are able to recite part of the count list (e.g., number words *one* through *five*) but without attaching any meaning to the count list. Children at this stage are unable to reliably construct sets for any number word and for this reason they are often referred to as non-knowers or pre-knowers. Following this stage, children develop an exact meaning for the number word *one* and these children can consistently construct sets of 1 object when instructed to do so, which is why they are classified as one-knowers. However, these children can't reliably give the correct sets for other numbers. Then, children learn the meaning of *two*; they can construct sets of one or two objects when asked for *one* and *two*, but can't construct the correct sets for other numerals. At this stage, children are called "two-knowers". Following the same pattern, children become "three-knowers" and sometimes "four-knowers". Then, after going through these "subset" stages, children's pattern of behavior at Give-N changes and they seem to be able to use the counting procedures to construct sets of larger cardinality under request. Children at that stage are often called Cardinal Principle or Counting Principle knowers (CP-knowers).

Some studies have investigated this sequence of acquisition in bilingual children and have shown that when bilingual children are classified as subset-knowers, they most often have different knower levels across their two languages (Sarnecka et al., 2021; Wagner et al., 2015). For example, a child could be classified as a two-knower in French and a three-knower in English. However, an

issue with those studies comes from the fact that they don't take into consideration the test-retest reliability of Give-N in their interpretation of the data. This is important because without knowing what the reliability of the task is, it becomes difficult to attribute the sources of cross-linguistic differences to reliability issues or to true differences in children's mappings between number words and cardinalities across languages. In this thesis, in addition to exploring the role of language in the development of numerical estimation in bilingual children, I also investigate the reliability of the Give-N task, a task commonly used to measure children's associative mappings between number words and cardinalities.

In contrast to Associative Mapping, Structure Mapping (or Analogical Mapping) is not based on a direct association between any specific number label and representation of magnitude but is instead based on a global analogy between the linguistic and nonlinguistic systems. Proponents of this system argue that the linguistic system (i.e., the count list) and the nonlinguistic systems (internal representations of magnitudes in the ANS) become connected as a whole to one another based on similarities in their structure (Sullivan & Barner, 2013, 2014). The structures of these systems are indeed similar in two important ways. First, both structures are ordered, that is, number words further along the counting list refer to larger quantities similar to how our internal representations of magnitudes increase as actual quantities perceived in the world increase. Second, both structures encode stable distance relations. For example, the number word *forty* comes twice as far along in the count list as *twenty* does, and *eighty* comes twice as far along as *forty*. Similarly, past studies have shown that these doubling relations are easily discriminated by the ANS (ratio of 1:2) and perceived as doubles, even in infancy (Xu & Spelke, 2000). Evidence for Structure Mapping comes from feedback effects (Sullivan & Barner, 2013). Specifically, studies have shown that participants' estimates can shift dramatically based on the feedback (or

anchor points) provided to them. For example, subjects adapt all their estimates whether they are told that the same set contains 100 dots or 1000 dots, regardless of how many dots the set actually contains. This provides evidence for a Structure Mapping between the linguistic and nonlinguistic systems because if number words were related to specific cardinalities via associative mapping, these feedback effects should not occur, such that estimates would change locally rather than globally.

Current thesis: Using bilingualism to study number concept development

One way of studying the role of language in the development of numerical concepts is to look for the different factors that drive changes in estimation abilities. There are a few potential factors that influence the development of estimation that can be inferred from the literature (Lipton & Spelke, 2005; Sullivan & Barner, 2013, 2014). One such factor is the maturation, in terms of acuity, of our nonlinguistic numerical systems (ANS and OTS). Other potential factors include language experience, such as having a strong knowledge of number words and the count list as well as education. Studying the relative contribution of these factors in monolingual children is limited because all these factors develop in parallel, making it difficult to isolate the relative role of each. In contrast, studying bilingual children provides a fruitful approach to investigating the role of language in numerical abilities because all the nonlinguistic factors are shared across languages within an individual child. For example, a French-German bilingual child may have fairly advanced knowledge of the count list in French but much less exposure to their counting system in German. Any differences in how this child estimates sets might be explained by differences in their linguistic abilities across these languages but would be difficult to explain via factors like ANS acuity, since these would not differ across languages.

In this dissertation, I explore how bilingual 3- to 7-year-old children estimate large and small cardinalities across their two languages as a case study to elucidate how language-specific experiences drive changes in number concepts. Specifically, I rely on verbal numerical estimation tasks in which participants are presented with flashed arrays of dots and are asked to estimate their numerosities. I will argue that individual differences in how children make numerical estimates are not purely due to changes to nonlinguistic factors such as the maturation of the ANS or executive function but are also due to changes to the linguistic system itself. I will also argue that language-specific experience starts influencing numerical representations across languages as soon as children learn their first number words and this effect of language is persistent during development.

In Chapter 1, I explore how French-English bilingual children estimate large sets of dots across their two languages and I present evidence that their estimates differ between languages. I then ask whether these differences in estimation are explained by differences in children's ability to access number words across languages. I find evidence that this is not the case. Instead, the data suggests that the differences in estimation are driven by differences in children's knowledge of the structure of the count lists across languages. In other words, estimation differences are not explained by the number words children are able to access across languages but by differences in how the number words that they can access are mapped onto nonlinguistic representations of magnitude. Overall, this provides evidence that changes in children's estimates are not only driven by maturation of their ANS acuity or access to number words but also by language-specific knowledge about the structure of the count list.

In Chapter 2, I address an important methodological concern when testing the bilingual population; the role of test-retest reliability. In particular, most studies conducted with bilinguals

assume that any difference observed between languages (e.g., number word understanding) is the result of true disparity in knowledge across languages, rather than being an artifact of test-retest reliability issues with the tasks used. In Chapter 2, I challenge this assumption and assess the reliability of the Give-a-Number task, a task that has been used in the bilingual literature to show early differences in the acquisition of number words. Specifically, I show that some of the differences found in knowledge of small number words across bilinguals' languages can be explained by reliability issues in the task. However, I replicate the finding that the understanding of the counting procedure transfers across languages in bilinguals; that is, children who are classified as CP-knowers in one language tend to be classified as CP-knowers in their other language as well.

Finally, whereas in Chapter 1 I tested the estimation of large sets, in Chapter 3, I asked when differences first emerge in estimation abilities by testing how children estimate very small numbers. Specifically, I use a subitizing task to test whether bilingual children have different mappings of small number words to small sets across languages. I present evidence that despite being proficient counters in both languages, bilingual children still show differences in their knowledge of small number words across languages. This suggests that learning the mappings between small number words and their cardinalities relies on language-specific experiences with the count list. It also raises the possibility that learning the counting procedure and acquiring meanings for small number words might follow different developmental trajectories. Taken together, these studies indicate that language-specific knowledge plays a fundamental role in how children come to associate the linguistic and nonlinguistic numerical systems and consequently, how they interpret the numerical properties of the word around them.

References

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, *86*(3), 201-221.

Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. *Child development*, *79*(4), 1016-1031.

Burr, D. C., Turi, M., & Anobile, G. (2010). Subitizing but not estimation of numerosity requires attentional resources. *Journal of Vision*, *10*(6), 20-20.

Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, *133*(1), 59-68.

Carey, S., & Barner, D. (2019). Ontogenetic origins of human integer representations. *Trends in cognitive sciences*, *23*(10), 823-835.

Dehaene, S. (2001). Précis of the number sense. *Mind & language*, *16*(1), 16-36.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. OUP USA.

Dehaene, S., & Changeux, J. P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of cognitive neuroscience*, *5*(4), 390-407.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Japel, C. (2008). " School readiness and later achievement": Correction to Duncan et al.(2007).

Feigenson, L., & Carey, S. (2003). Tracking individuals via object-files: evidence from infants' manual search. *Developmental Science*, *6*(5), 568-584.

Feigenson, L., & Carey, S. (2005). On the limits of infants' quantification of small object arrays. *Cognition*, *97*(3), 295-313.

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychological science*, *13*(2), 150-156.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*(7), 307-314.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, *306*(5695), 496-499.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the" Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental psychology*, *44*(5), 1457.

Ifrah, G. (2000). The universal history of numbers: From prehistory to the invention of the computer, translated by David Vellos, EF Harding, Sophie Wood and Ian Monk.

Indow, T., & Ida, M. (1977). Scaling of dot numerosity. *Perception & Psychophysics*, *22*(3), 265-276.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: kindergarten number competence and later mathematics outcomes. *Developmental psychology*, *45*(3), 850.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395-438.

Lipton, J. S., & Spelke, E. S. (2003). Origins of number sense: Large-number discrimination in human infants. *Psychological science*, *14*(5), 396-401.

Lipton, J. S., & Spelke, E. S. (2004). Discrimination of large and small numerosities by human infants. *Infancy*, *5*(3), 271-290.

Lipton, J. S., & Spelke, E. S. (2005). Preschool children's mapping of number words to nonsymbolic numerosities. *Child development*, *76*(5), 978-988.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of experimental psychology: animal behavior processes*, *9*(3), 320.

Menninger, K. (1969). *Number words and number symbols: A cultural history of numbers.* Courier Corporation.

Piazza, M. (2011). Neurocognitive start-up tools for symbolic number representations. *Space, time and number in the brain*, 267-285.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, *306*(5695), 499-503.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation?. *Psychological science*, *19*(6), 607-614.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*(3), 662-674.

Sarnecka, B. W., & Lee, M. D. (2009). Levels of number knowledge during early childhood. *Journal of experimental child psychology*, 103(3), 325-337.

Sarnecka, B. W., Negen, J., Scalise, N. R., Goldman, M., & Rouder, J. (2021). The real preschoolers of Orange County: Early number learning in a diverse group of children.

Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental science*, *11*(5), 655-661.

Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers' numerical understanding. *Journal of educational psychology*, *101*(3), 545.

Spaepen, E., Coppola, M., Flaherty, M., Spelke, E., & Goldin-Meadow, S. (2013). Generating a lexicon without a language model: Do words for number count?. *Journal of Memory and Language*, *69*(4), 496-505.

Spelke, E. S. (2017). Core knowledge, language, and number. *Language Learning and Development*, *13*(2), 147-170.

Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition*, *78*(1), 45-88.

Sullivan, J., & Barner, D. (2013). How are number words mapped to approximate magnitudes?. *The Quarterly Journal of Experimental Psychology*, *66*(2), 389-402.

Sullivan, J., & Barner, D. (2014). Inference and association in children's early numerical estimation. *Child development*, *85*(4), 1740-1755.

Wagner, K., Kimura, K., Cheung, P., & Barner, D. (2015). Why is number word learning hard? Evidence from bilingual learners. *Cognitive psychology*, *83*, 1-21.

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological science*, *10*(2), 130-137.

Wilkey, E. D., & Ansari, D. (2020). Challenging the neurobiological link between number sense and symbolic numerical abilities. *Annals of the New York Academy of Sciences*, *1464*(1), 76-98.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155-193.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive psychology*, *24*(2), 220-251.

Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition*, *89*(1), B15-B25.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1-B11.

CHAPTER 1

Language-specific numerical estimation in bilingual children

Elisabeth Marchand[1], Shirlene Wade[2,3], Jessica Sullivan[4] and David Barner[1]

[1]Department of Psychology, University of California, San Diego

[2]Department of Brain and Cognitive Sciences, University of Rochester

[3]Department of Psychology, University of California, Berkeley,

[4]Department of Psychology, Skidmore College

Abstract

We tested 5- to 7-year-old bilingual learners of French and English ($N = 91$) to investigate how language-specific knowledge of verbal numerals affects numerical estimation. Participants made verbal estimates for rapidly presented random dot arrays in each of their two languages. Estimation accuracy differed across children's two languages, an effect that remained when controlling for children's familiarity with number words across their two languages. In addition, children's estimates were equivalently well ordered in their two languages, suggesting that differences in accuracy were due to how children represented the relative distance between number words in each language. Overall, these results suggest that bilingual children have different mappings between their verbal and nonverbal counting systems across their two languages and that those differences in mappings are likely driven by an asymmetry in their knowledge of the structure of the count list across their languages. Implications for bilingual math education are discussed.

Introduction

The human ability to encode number in natural language has important consequences for our species, allowing us to perform exact numerical computations, construct models of the physical world, and make precise predictions regarding distant places and hypothetical situations. Human cultures that lack labels for quantities greater than three or four are generally unable to represent larger sets exactly (e.g., Coppola, Spaepen, & Goldin-Meadow, 2013; Dixon, 2004; Epps, 2006; Frank, Everett, Fedorenko, & Gibson, 2008; Gordon, 2004; Pica, Lemer, Izard, & Dehaene, 2004; Spaepen, Coppola, Flaherty, Spelke, & Goldin-Meadow, 2013; Spaepen, Coppola, Spelke, Carey, & Goldin-Meadow, 2011). Similarly, young children who have not yet learned to count are restricted to representing and comparing larger sets approximately (e.g., Feigenson, Dehaene, & Spelke, 2004; Wynn, 1990, 1992). Although body count systems (Ifrah, 2000; Menninger, 1969), ancient counting boards and abaci (Frank & Barner, 2012; Hatano, Miyake, & Binks, 1977) and nonverbal cognitive systems like the evolutionarily ancient approximate number system (ANS) (Dehaene, 1997) all provide ways for humans to encode and reason about numerosity, language affords humans a distinctly flexible and productive format for describing and manipulating number and, by some accounts, may be the conceptual basis for the creation of other symbolic representations of number (e.g., Chomsky, 2008; Di Sciullo, 2012; Spelke, 2017; Watanabe, 2017).

Like many nonhuman animals, humans are able to represent the approximate magnitudes of large sets independent of language using the ANS. For example, human infants, pigeons, rats, and even fish can discriminate different quantities of objects on the basis of their numerical ratio, consistent with Weber's law (Agrillo, Dadda, Serena, & Bisazza, 2008; Platt & Johnson, 1971; Scarf, Hayne, & Colombo, 2011; Xu & Spelke, 2000). The acuity of these representations

improves quickly during infancy (Xu & Spelke, 2000; Xu, Spelke, & Goddard, 2005) and continues to change over the first 20 years of life into adulthood (Barth, Kanwisher, & Spelke, 2003; Halberda & Feigenson, 2008). Still, even in adults, many small exact differences that we can represent in language, such as 87 versus 88, cannot be discriminated by the ANS, suggesting that this system alone cannot explain the exact nature of linguistic numerical representations. Such findings raise the question of how language might allow humans to go beyond the limits of the ANS and represent large exact magnitudes.

Several recent studies have argued that exact number representations (e.g., number word meanings, addition facts, multiplication tables) are both language dependent (meaning that they do not occur in the absence of language) and language specific (meaning that, in bilinguals, representations in one language are not automatically available to other languages), whereas approximate representations and computations are independent of natural language, such that training in one language generalizes automatically to a second language (L2). For example, Spelke and Tsivkin (2001) trained Russian–English bilingual adults in one of their two languages on four types of mathematical problems, two of which required exact numerical responses (e.g., addition of multidigit numbers) and two of which required approximate solutions (e.g., computing cube roots). After training, participants responded more slowly in their untrained language to questions that required precise answers (indicating language specificity), whereas there was no effect of language for questions that required approximate answers (for similar results, see Saalbach, Eckstein, Andri, Hobi, & Grabner, 2013). Similarly, in a neuroimaging study of bilingual adults, Dehaene, Spelke, Pinel, Stanescu, and Tsivkin (1999) found that problems requiring exact computations, such as multiplication tables, recruited areas of the brain involved in word association processes, whereas approximate number tasks, such as summing rapidly displayed

arrays of dots, relied on areas of the brain associated with visuospatial processing rather than language, suggesting language dependence (for similar results in monolinguals, see Dagenbach & McCloskey, 1992; Dehaene & Cohen, 1997; Dehaene, Molko, Cohen, & Wilson, 2004; Delazer & Benke, 1997; Domahs & Delazer, 2005; Ischebeck et al., 2006; Lampl, Eshel, Gilad, & Sarova-Pinhas, 1994; Lee, 2000; Lemer, Dehaene, Spelke, & Cohen, 2003; Pesenti, Seron, & van der Linden, 1994; van Harskamp & Cipolotti, 2001; van Harskamp, Rudge, & Cipolotti, 2002).

Although it is clear that approximate and exact linguistic representations of number arise from distinct learning processes and rely on different cortical regions of the brain, previous studies leave open the mechanisms by which the systems become related, and to what degree mappings between systems might differ across a bilingual learner's languages. In order to estimate the number of dots presented on a screen — an ability that emerges during development sometime after 4 years of age — children must relate symbolic representations of number (i.e., words in their verbal count list) to non-symbolic ANS values (Barth, Starr, & Sullivan, 2009; Gunderson, Spaepen, & Levine, 2015; Le Corre & Carey, 2007; Lipton & Spelke, 2005; Mejias & Schiltz, 2013; Siegler & Booth, 2004; Sullivan & Barner, 2014; Wagner & Johnson, 2011). Whereas much is known about changes in children's estimation abilities over development, less is known about the mechanisms that drive change. Changes in estimation ability could be driven by maturation of the ANS and/or acquired experience with the count list. Here, we asked how experience with a count list influences estimation performance, controlling for ANS ability. We approached this question by studying bilingual children who were acquiring two count lists (French and English). One possibility is that when children acquire two count lists (e.g., one in French and another in English), knowledge of these lists fails to transfer across languages because the words — which differ phonologically across languages — are mapped on an item-by-item basis to individual ANS

values, such that knowledge cannot generalize to another language. Another possibility, not incompatible with the first one, is that estimation does not depend solely on item-based associations between words and ANS values, but also draws on more general knowledge that children may glean from experience in estimating, or from the structure of counting itself (Izard & Dehaene, 2008; Sullivan & Barner, 2014). This approach allows us to investigate the independent effect of language (e.g., familiarity with number words, the structure of the count list) on estimation.

Although no previous study has tested these questions in the context of estimation, evidence from young bilingual children indicates that knowledge of the first number words (i.e., meanings for ‘‘one,” ‘‘two,” and ‘‘three”) does not transfer across languages, probably because these words are represented as item-based associations between words and cardinal values (Wagner, Kimura, Cheung, & Barner, 2015). Critically, however, the same study found that the knowledge of counting procedures that allow children to count and accurately give large sets (sometimes called ‘‘cardinal principle” knowledge) does transfer across a bilingual child’s two languages. This raises the possibility that if early estimation abilities are governed chiefly by item-based associations between words and ANS values, transfer might not occur between languages. However, if estimation is governed in part by more general principles — not specific to particular words — transfer might occur. For example, a child who has extensive experience in making estimates in English might notice a ‘‘later – greater” principle — that numbers later in the count list denote greater quantities (Davidson, Eng, & Barner, 2012; Le Corre, 2014) and that the relative distance between two magnitudes corresponds to the distance between their verbal labels in the count list. This type of general principle might be learned in one language and extended to an L2, thereby facilitating learning in the L2.

Understanding how estimation abilities are mediated by language in bilingual learners not only is theoretically important but also has practical implications. First, given the global prevalence of exposure to bilingual education, it is important to know whether numerical knowledge acquired in one language will readily transfer to a child's L2 or whether educators should dedicate time to conducting training in both languages separately. For example, in the context of a bilingual immersion curriculum, children who live in a culture that requires math fluency in one language (e.g., English) but who receive their primary training in a different language (e.g., French) may benefit from learning certain foundational math skills in both of their languages. Second, previous studies have found relationships between estimation abilities and symbolic arithmetic (Booth & Siegler, 2008; Desoete, Ceulemans, De Weerdt, & Pieters, 2012; De Smedt, Noël, Gilmore, & Ansari, 2013; Fazio, Bailey, Thompson, & Siegler, 2014; Holloway & Ansari, 2009; Sullivan, Frank, & Barner, 2016). Bilingual learners provide an interesting test case because individual learners make estimates in multiple languages and yet draw on a single non-symbolic ANS to do so. Consequently, studying bilingual learners permits the isolation of effects of language experience on estimation — separate from, for example, the acuity of ANS representations — and which of these components might better explain later mathematics achievement.

To explore the role of language in the development of estimation, we presented 5- to 7-year-old French-English bilinguals with a dot array estimation task in their two languages. We also tested how high children could count in each language as a proxy of relative exposure to each language's counting system. We then asked whether children generate different estimates in their two languages in terms of (a) accuracy and (b) degree of variability (i.e., uncertainty). We also asked whether any such differences might be explained by children's knowledge of counting structure in each language, including factors such as their access to number words in each language

and their knowledge of the later-greater principle — that is, whether on consecutive trials an increase or decrease in the size of a presented dot array was reflected by a verbal estimate that was earlier or later in the count list.

Method

Participants

In total, 93 French-English bilingual 5- to 7-year-old children participated in the study (M = 6.8 years, SD = 0.8, range = 4.9 – 8.0). This sample size represents the maximum number of children who could be tested during a 3-week field trip to Comox, British Columbia, Canada (n = 64), plus a smaller group of children tested at a French immersion school in San Diego, California (n = 26), or at a lab at the University of California, San Diego (UCSD) (n = 3), in the southwestern United States. Two children were excluded due to failure to complete more than half of the trials, leaving a final total sample of 91.

The field site in British Columbia was chosen because of the availability of both first-language (L1) English-speaking and L1 French-speaking children who lived in an English-dominant environment but attended French immersion schools, resulting in high levels of balanced bilingual number knowledge. At one school, attendance was restricted to children whose caregivers spoke French as an L1 (n = 34), whereas the other school targeted children from English-speaking households (n = 30). The school in San Diego (n = 26) served children from a variety of linguistic and cultural backgrounds (e.g., French citizens, French Canadians, monolingual English families interested in French education). In all three schools, formal instruction was given almost exclusively in French, including math curriculum. As part of the consent process, parents received a short questionnaire asking whether their children were

bilingual. Children who were not identified as bilingual by caregivers did not participate in the study. Participants were, in general, from families of medium to high socioeconomic status. The study received approval by the ethics committee of UCSD.

Procedure

Each participant completed a dot array estimation task and a counting assessment in both English and French. Each child received two blocks of tasks — an English block and a French block — the order of which varied between participants. Within each block, children were always presented with the dot array estimation task before the highest count task. In the English block, children were tested by an English-speaking researcher. In the French block, children were tested by a French-English bilingual researcher. The majority of participants (n = 75) were tested in both languages by the same French-English bilingual researcher, and the remainder (n = 16) were tested by two different researchers to maximize data collection in the available time. Testing lasted on average 30 min. All participants were tested at their respective schools, with the exception of 3 children who were tested at the UCSD Language and Development Lab.

Dot Array Estimation Task

Participants were presented with dot arrays on a computer screen. Arrays ranged in magnitude from 4 to 98. Some items were presented eight times each (4, 16, 32, 60, and 80), and to avoid boredom these trials were interspersed with other items that were presented twice each (8, 12, 24, 44, 70, and 98). Stimuli were presented to each participant in two blocks of trials, where Block A and Block B contained identical items but in different randomized orders. Each participant received both blocks in both languages, and the order of the blocks was counterbalanced. Hence,

participants were assigned to one of four conditions in a 2 (English first vs. French first) x 2 (Block A first vs. Block B first) design.

Participants were first introduced to an experimenter who explained that they would be tested in two languages. This sentence was presented in French for children tested with the French block first and in English for children tested with the English block first. Next, children were introduced to the task in that block's language (e.g., in the French block, instructions were in French). Children were instructed to look at the dots on the screen and guess how many there were. All participants' responses were elicited verbally in the language of test (French or English depending on language block) and were recorded by the experimenter. The experimenter provided no information to children about the range of numbers in the task. Noninformative verbal encouragement was given to the participants to keep them motivated (e.g., ‘‘You are very attentive", ‘‘You already did half of the game!"). Within each language block, participants were tested for 10 min or until the completion of all 52 trials, whichever came first.

Highest Count Task

Participants were asked to count as high as they could in English and in French. They were stopped at 100 or when they made their first error.

Results

Data Management

Highest count

We recorded each child's highest count in each of the two languages, which was defined as the highest number counted to without making an error (with a maximum value of 100), in order to describe the counting abilities of our sample.

Estimation

Before conducting our primary analyses, we excluded all estimates that were not provided in the format of a unique verbal numeral (n = 64/9464) such as expressions of addition or multiplication (e.g., ''There are six plus three!") and other non-numbers (e.g., ''Eleventy billions"). We also removed outlier estimates that were 10 times larger or smaller than the presented numerosities (n = 247/9464) or that were at least 5 standard deviations from the mean estimation for each numerosity presented (n = 35/9153). Except for our ordinality analyses, all analyses reported below were performed on the remaining 9118 responses. Ordinality analyses cannot be affected by outliers (because they consider only order and not absolute distance between estimates); therefore, outliers were not removed for this analysis.

We computed four main measures of estimation performance: raw estimates, percentage absolute error (PAE), coefficient of variation (COV), and ordinality. Raw estimates were children's numerical estimates. Consistent with previous work (Barth et al., 2009; Lipton & Spelke, 2005; Sullivan & Barner, 2014), our first set of analyses tested the relation between children's verbal estimates and the number of dots presented. To determine whether language affected estimates, analyses reported probed whether language predicted a difference in the slope of estimates (i.e., whether there was an interaction between language and numerosity).

The remaining three analysis types were planned to probe the nature of any effects of language on estimation. First, because it is possible for estimates to exhibit a slope approaching 1

despite being inaccurate (e.g., in cases where the intercept is not 0, in cases where estimates for a particular numerosity were highly variable but averaged out to the ‘‘correct” response), our second set of analyses measured accuracy as the PAE of estimates, defined as the absolute distance between an estimate and its target numerosity on a particular trial:

$$\text{PAE} = \left|\frac{\text{Estimate} - \text{Numerosity}}{\text{Numerosity}}\right| \times 100$$

PAE values were log transformed to normalize a skewed distribution of residuals (to satisfy the assumptions of our linear regression analyses; see below). Second, we also asked how variable estimates were by calculating the COV, that is, the standard deviation of estimates relative to the mean estimate for a particular target:

$$\text{COV}_N = \frac{\text{Standard Deviation of Estimate}_N}{\text{Mean Estimate}_N}$$

For each participant, this generated one COV per numerosity (*N*) per language.

Finally, we calculated the ordinality of participants’ responses (as a measure of later-greater knowledge). Participants’ raw estimates, COV, and PAE all are sensitive to the relative distance between their estimates. For example, if a participant sees a set of 40 dots and estimates ‘‘forty,” the child’s PAE and raw estimate will be different than if the child had estimated ‘‘forty-one” (and, in most contexts, so will the child’s COV). However, as noted by Sullivan and Barner (2013, 2014), children often make estimates that are wildly inaccurate despite being well ordered; for example, if a child estimated ‘‘four” after seeing 4 dots and then saw 44 dots on the next trial, the child’s response on this second trial would be considered ordinal whether the child responded ‘‘five,” ‘‘five million,” or any other number greater than 4. Testing for ordinality is important

because the presence of ordinal estimates would indicate that poor estimation accuracy cannot be attributed to poor knowledge of how number words are ordered or that numbers later in the count list denote greater quantities. Instead, such evidence would implicate other factors such as the problem of relating distance between numbers in the count list to corresponding differences between ANS values. We return to the significance of this distinction in the Discussion.

To calculate the ordinality of each numerical estimate, *n*, we coded whether it differed from the previous trial's estimate in the correct direction. For example, if a child saw 4 dots on trial *n* - 1 and then saw 44 dots on trial *n*, the child's estimate was coded as ordinal only if the child provided a larger estimate on trial *n* than on trial *n* - 1.

Preliminary analyses

For these and all analyses, models were generalized linear mixed-effects models constructed in R using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2019). All model outputs are available in our Open Science Framework (OSF) repository (https://osf.io/f8v7b/?view_only= 01910a3ad63c4fa39071617e24c10ab3).

Highest Count

Highest count was used to measure children's familiarity with the count lists in French and in English. As shown in Figure 1.1, participants generally counted higher in English than in French ($t(90) = 5.59$, $p < .0001$). On average, children counted up to 50.5 in French ($SD = 31$, range = 3 – 100) and up to 71 in English ($SD = 37$, range = 8 – 100). In total, 50 children counted to 100 in English, whereas 14 children counted to 100 in French. In addition, 49 children counted higher in English than in French, and 17 children counted higher in French than in English. A total of 25

children were ‘‘balanced counters” in that their highest count differed by less than 10% across their two languages.

Estimation

We conducted preliminary analyses to ensure that there were no effects of language order (i.e., which language children were tested in first) on estimates. To test this, we predicted each measure of estimation performance (raw estimates, ordinality, PAE, and COV) from language order and numerosity; for all models, participant and numerosity were treated as random factors[1]; we generated *p* values by comparing our full model with one that excluded language order. We found no effect of language order for any of our measures of estimation (all *ps* > .10).

Because there are typically age effects on estimation performance, we also conducted preliminary analyses predicting each estimation measure from age (*z* scored), numerosity, and their interaction. For raw estimates, there was a significant interaction of age and numerosity ($B = .12$, $SE = .01$, $t = 14.28$, $p < .0001$). There were significant effects of age for PAE ($B = .25$, $SE = .03$, $t = 8.99$, $p < .0001$) and ordinality ($B = .35$, $SE = .05$, $z = 7.04$, $p < .0001$). For COV, there were no effects of age ($B = .02$, $SE = .01$, $t = 1.31$, $p > .10$) or numerosity ($B = .001$, $SE = .002$, $t = 0.93$, $p > .10$).

Because language order had no effect in preliminary analyses, we did not consider it in subsequent analyses. However, we included age, numerosity, and their interaction in our raw estimate models. We included age and numerosity in our PAE models and ordinality models. We

[1] Numerosity was also added as a random factor in our models to account for the fact that some estimation items were presented eight times, whereas others were presented only four times (see Method).

did not include age or numerosity in our COV models because it had no effect in preliminary analyses.

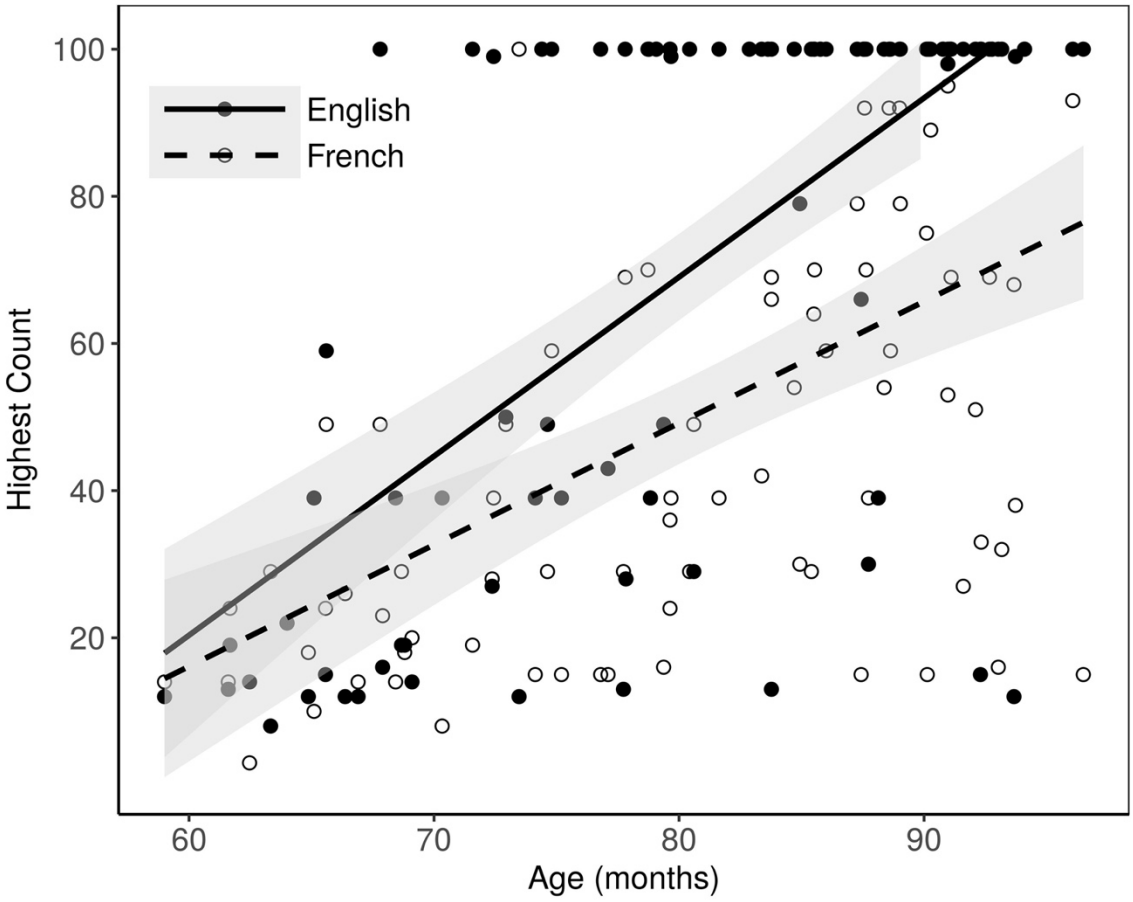

Figure 1.1: Highest count performance in English and French by age.

Note. Highest count performance in English (black/solid/closed circles) and French (white/dotted/open circles) by age.

Main Analyses

All analyses reported below were conducted using mixed-effects models, such that all estimates were entered into a single regression, with slopes fitted for each participant.

Our primary question was whether language of test affected estimation performance. To test this, we first constructed a model predicting raw estimates from age (in months), numerosity, language of test, and their interactions. If estimates differ across languages, we should find an interaction of numerosity and language of test; this would indicate that there is a difference in the slope of estimates for English relative to for French. In our main analyses, we planned to interpret the highest order interaction containing language of test. We found a significant three-way interaction among age, numerosity, and language of test ($B = .11$, $SE = .01$, $t = 11.02$, $p < .0001$) (see Figure 1.2), which reflected a difference in estimation behavior between French and English languages that was greatest among younger children.

This first finding suggests that fluent bilingual children, who are immersed through schooling and their community in a French- and English-speaking environment, exhibited significantly different estimation abilities across their two languages. However, this first analysis leaves open the precise nature of this difference. A strong interpretation of our finding is that children's estimates differ across their languages because children have made different mappings between number words and the ANS in each case. However, although our analysis found a difference in estimates across languages, it leaves open the possibility that differences were due to the fact that children accessed and used different number words in each of their languages. For example, a child who is familiar with numbers up to approximately 80 in one language and up to approximately 150 in the other language might make similar use of the numbers up to 80 in each language — perhaps with equal accuracy or noise — but perform differently for larger numbers that the child can access in only one language (e.g., more or less accurately).

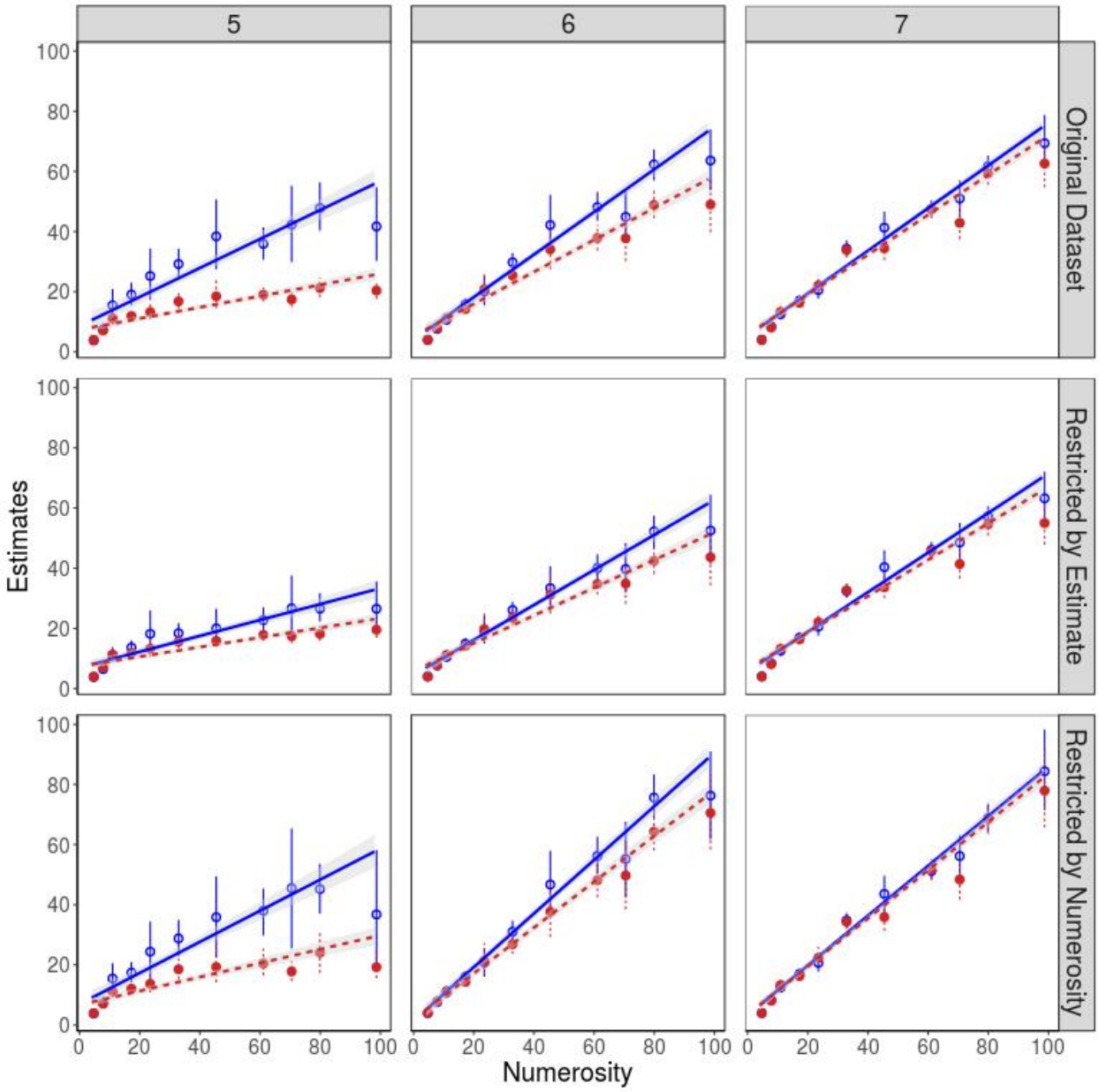

Figure 1.2: Relationship between numerosity and estimates in English and French.

Note. Relationship between numerosity and estimates in English (solid/open circles) and French (dashed/closed circles). Panels are grouped by age in years (5-year-olds: n = 20; 6-year-olds: n = 30; 7-year-olds: n = 41). Data points represent binned means. Bars denote bootstrapped 95% confidence intervals.

To identify whether children made different use of the numbers that overlapped across their two languages, we first identified each child's ''highest common estimate," which was defined as the highest number that the child produced in *both languages* during the estimation task (e.g., if

the child's highest verbal estimate in English was 80 and in French was 150, that child's highest common estimate was 80). We reasoned that if access to number words limited performance and caused cross-linguistic differences in estimation, it should have done so only outside of this productive estimation range. Differences within the highest common estimate, however, would suggest that children's estimates differ not only because of their differing abilities to access number words, but also because of how the words they can access are mapped to ANS values. In a first post hoc analysis[2] that included only verbal estimates that were within the child's highest common estimate range, we once again found a three-way interaction among numerosity, language of test, and age (B = .03, SE = .01, t = 3.39, p < .0001). Our second post hoc analysis considered only trials where the presented numerosity had a value that was within the child's highest common estimate (e.g., considering only those trials on which the target numerosity was 80 or smaller). Once more, we found a three-way interaction among numerosity, language of test, and age (B = .09, SE = .01, t = 7.56, p < .0001). Thus, even when considering only trials that were within the child's estimation range for both languages, estimates differed as a function of language, suggesting that bilingual children appear to have different mappings between ANS values and familiar number words across their two languages. Our next analyses addressed the nature of this difference.

As a first step to understanding the nature of children's different estimates across French and English, we considered estimation accuracy, as measured by the PAE of estimates (i.e., log PAE). If language affects the accuracy of estimates, we should expect PAE to differ for English

[2] All 91 participants were represented in these post hoc analyses, which contained 76.34% to 92.94% of the original dataset (depending on the analysis). None of the analyses on raw estimates, PAE, and ordinality was conducted on datasets containing fewer than 7400 trials. The dataset used for the main analysis of COV contained 1992 trials because a unique COV was calculated for each numerosity in each language per child. The COV post hoc analyses were performed on datasets containing 1937 and 1634 trials.

versus French. Evidence for this would include either a main effect of language of test or an interaction of numerosity and language of test, such that participants are more accurate at estimating in one numerical range for one language but in a different numerical range for the other language. To test this, we predicted PAE from age, numerosity, language of test, and the interaction of numerosity and language of test. This analysis found an effect of age (B = .25, SE = .03, t = 9.00, $p < .0001$), reflecting the fact that accuracy was greater among older children (see Figure 1.3). Importantly, there was also a significant interaction of numerosity and language of test (B = .005, SE = .001, t = 7.29, $p < .0001$), suggesting that, at least in some numerical ranges, there was a difference in accuracy across the child's two languages. Following the method outlined above, we next asked whether the effect of language persisted when examining only numbers within the child's highest common estimate. We found that the interaction between numerosity and language of test remained even when considering only verbal estimates or magnitudes within the child's highest common estimate (trimmed estimation range: B = .004, SE = .001, t = 5.04, $p < .0001$; trimmed numerosity range: B = .004, SE = .001, t = 4.56, $p < .0001$) (see Figure 1.3). As shown in Figure 1.3, this result was likely driven by the underestimation of larger numbers (60–100 range) in French for 5- to 7-year-olds (see Figures 1.2 and 1.3). Thus, children's estimates in French and English differed in accuracy, and this difference remained when considering only numbers used in both languages.

As noted by an anonymous reviewer, although PAE is a standard measure used to capture estimation accuracy, it is also sensitive to variability around the mean of a participant's estimates. This is not a problem from the perspective of testing whether children make different estimates across their languages, but it may make it more difficult to isolate accuracy as distinct from variability. To address this, we reproduced our PAE analyses using a modified measure of

accuracy, signed error rate, defined as (estimate - numerosity)/numerosity. Using this measure, we again found that the accuracy of children's estimates differed reliably across their two languages. Specifically, we found an interaction between age and numerosity (B = .001, SE = .0002, t = 5.80, p < .0001) and a significant interaction of numerosity and language of test (B = .001, SE = .001, t = 2.81, p < .001). These effects remained significant when considering only trimmed data (trimmed estimation range: B = .001, SE = .0004, t = 2.25, p < .05. When considering trimmed numerosity range, the main effect of language of test remained significant (B = .72, SE = .008, t = 5.46, p < .0001), although the interaction between language of test and numerosity did not (p = .08). Thus, this series of post hoc analyses suggests that when accuracy is isolated from the variability of estimates (i.e., using signed error rate as a measure), our results resemble those reported for PAE, although in one case (trimmed numerosity range) the effect of language of test does not differ as a function of numerosity.

We next asked whether children's estimates differed in how variable they were across children's two languages — for example, whether, on average, the standard deviation of estimates for a particular numerosity was greater in one language than in the other language. We measured this by computing the COV for each target value, for each of children's two languages, as described above. Next, we constructed a model predicting COV from language of test (recall that neither numerosity nor age was related to COV in preliminary models and, therefore, was excluded from this model). Children's COV in English (M = .35) was modestly greater than that in French (M = .32), a potentially surprising result given that children were overall more accurate in English than in French (see Figure 1.4). However, this effect did not reach the alpha threshold of p < .05 (B = .023, SE = .0118, t = 1.96, p = .0502). When analyses were restricted to only estimates within children's productive estimation range, or when analyses were restricted to only numerosities

within children's productive estimation range, language of test did not predict COV (all *ps* > .10). This suggests that, to the extent that there was a difference in COV in the first analysis, this was likely due to greater variability for numbers that children produced only in English.

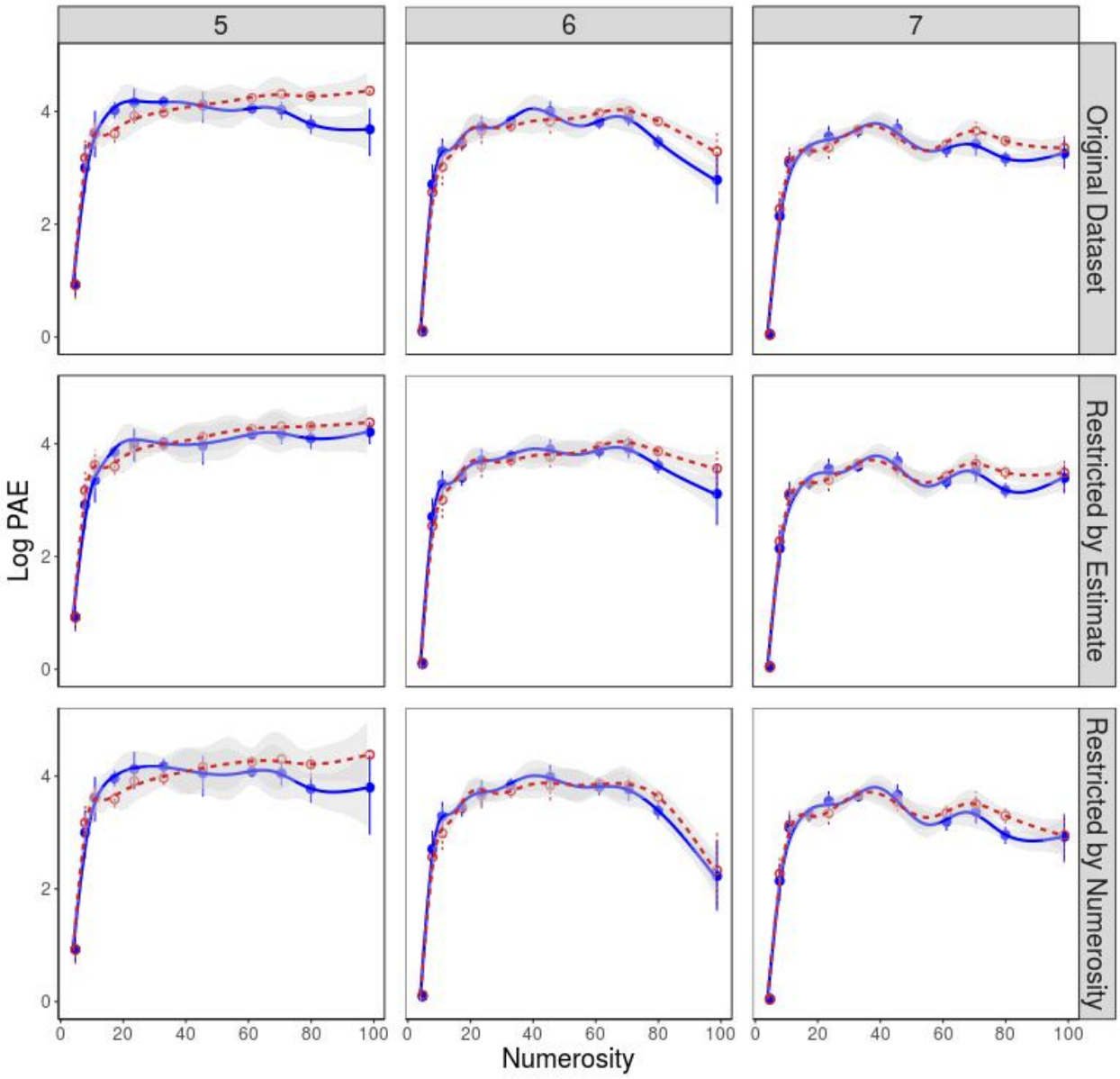

Figure 1.3: Relationship between numerosity and percentage absolute error in English and French

Note. Relationship between numerosity and log percentage absolute error (PAE) in English and French across age (5-year-olds: n = 20; 6-year-olds: n = 30; 7-year-olds: n = 41). Data points represent binned means. Error bars are bootstrapped 95% confidence intervals.

Our last set of analyses considered ordinality. Recall that the accuracy and variability of estimates each may be affected by (a) children's knowledge that words later in the count list denote larger magnitudes (i.e., ordinality/later-greater), (b) children's knowledge of how relative distance

between verbal estimates in the count list corresponds to distance between ANS values (numerical distance), or (c) both. To assess whether ordinality differed across languages, we constructed a model predicting ordinality from language of test, numerosity, and age. We found a main effect of age (B = .35, SE = .05, z = 7.05, p < .0001), such that older children gave more ordinal estimates than younger children, and a main effect of language of test (B = .14, SE = .06, z = 2.45, p = .014), such that ordinality rates were higher in English (82.8%) than in French (80.9%), although this effect was very small (2% difference) (see Figure 1.5). There was no reliable effect of numerosity (B = .02, SE = .01, z = 1.76, p = .079). When we considered only estimates that were within the child's highest common estimate, the difference between languages became smaller (English = 83.8%; French = 82.2%) and the effect of language of test was no longer significant (B = .107, SE = .062, z = 1.74, p = .083). Likewise, when we considered only numerosities that were within the child's productive estimation range, we also found no significant effect of language on ordinality (English = 83.8%; French = 82.4%; B = .107, SE = .064, z = 1.67, p = .095). These data indicate that although children exhibit significant differences in the ordinality across all their estimates in French and English, these differences are very small and cannot likely explain the much larger differences in accuracy reported above. In addition, when analyses are restricted to numbers that are familiar across both of a child's languages, these already small effects become nonsignificant.

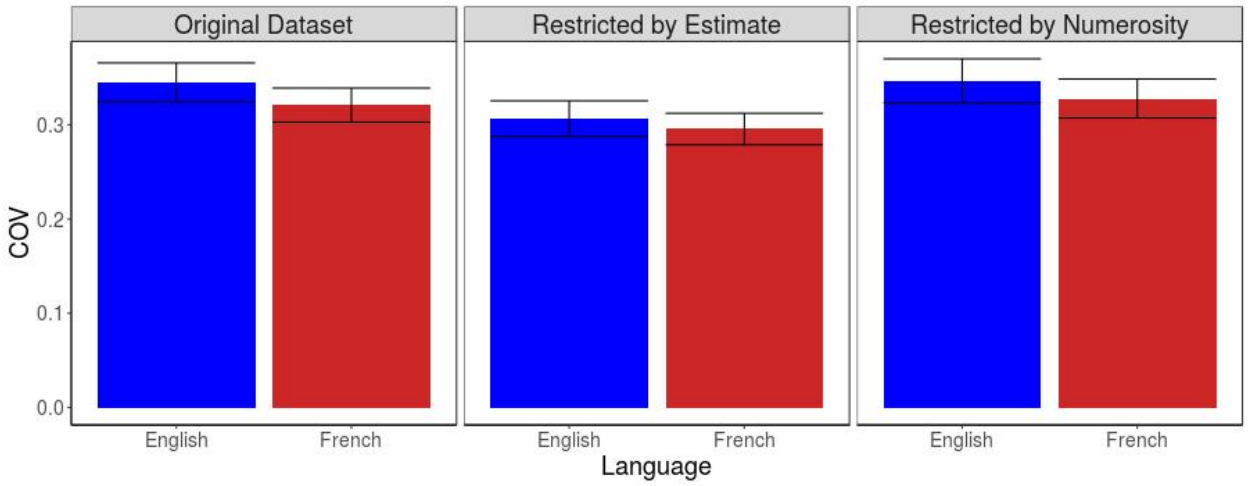

Figure 1.4: Average coefficient of variation in English and French.

Note. Average coefficient of variation (COV) in English (left bars) and French (right bars). Error bars represent 95% confidence intervals.

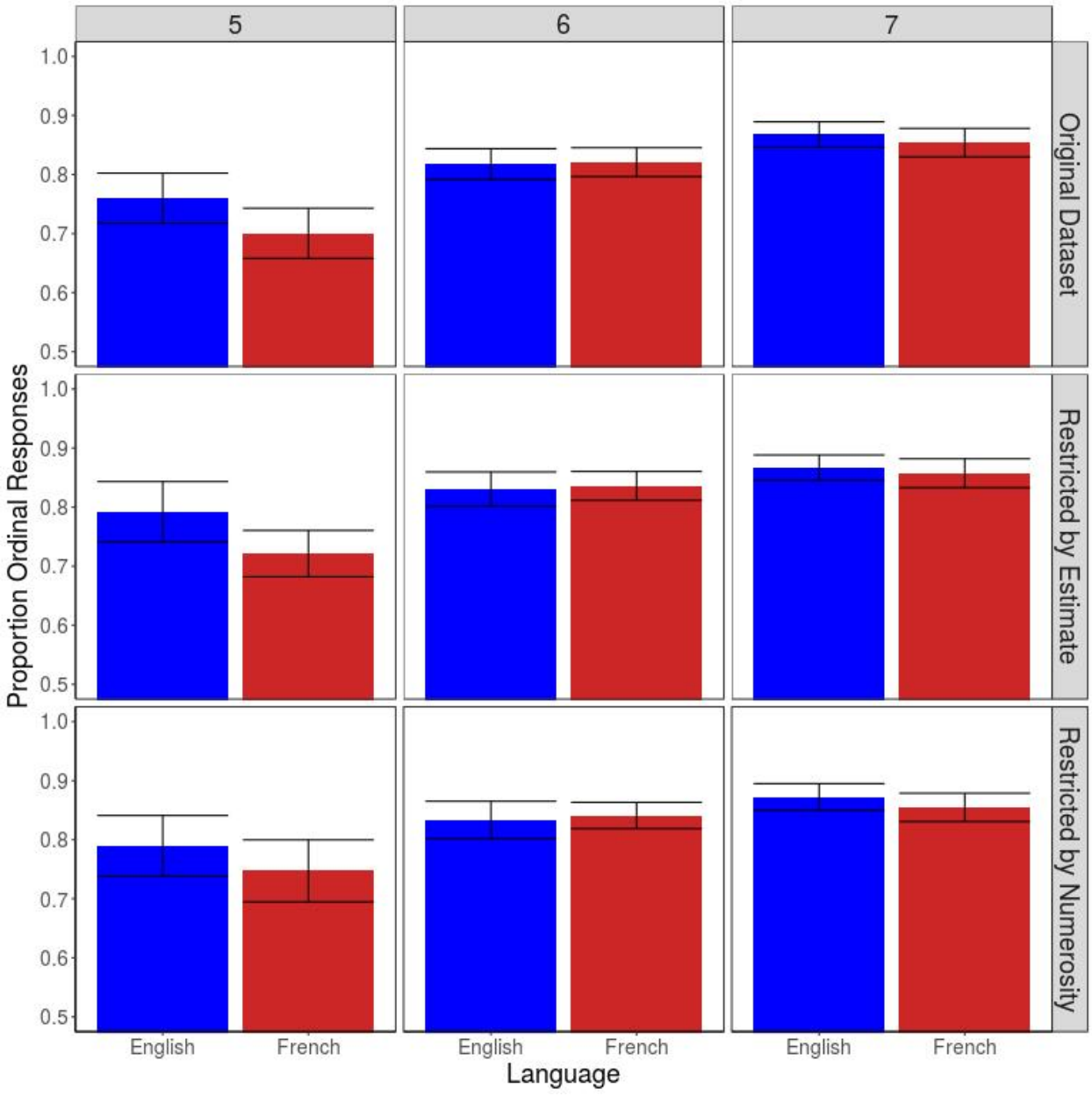

Figure 1.5: Average proportion of ordinal responses in English and French

Note. Average proportion of ordinal responses in English (left bars) and French (right bars) across age (5-year-olds: n = 20; 6- year-olds: n = 30; 7-year-olds: n = 41). Error bars represent 95% confidence intervals.

Discussion

We tested whether, when children acquire multiple languages, their estimation abilities differ across their two languages. In particular, we asked whether children's mappings between number words and magnitudes are language specific (e.g., in French vs. English) or whether certain components of these mappings might transfer across languages in bilingual learners. To do this, we asked French-English bilingual children to make dot array estimates in each of their two languages. By eliciting estimates in both languages, we examined the effect of language independent of the acuity of nonlinguistic (ANS-based) numerical representations, which remained constant (because language was manipulated within participants). We found that estimates differed across children's two languages and, in particular, that their estimates were less accurate in French than in English, especially among younger participants. Interestingly, this difference in estimation performance remained even when analyses were restricted to the range of numbers that children produced in their weaker language (i.e., their ''highest common estimate") as well as when analyses were restricted to dot array numerosities that fell within this same range. These data suggest that the accuracy of mappings between number words and numerical magnitudes differs across a child's two languages and that this difference cannot be explained by differences in the precision of the ANS (because children use the same ANS system in both languages). These data are most consistent with the view that the structures that support mature estimation are, at least in part, specific to particular languages and, therefore, do not readily transfer from a highly trained language to a less trained language.

How might a child's linguistic representations of number differ, such that estimates differ across languages? We addressed this question by asking which types of knowledge differ across languages and, therefore, are language specific and not readily transferred from one language to the other language. As noted above, we found that accuracy differed significantly across children's

French and English estimates. However, other aspects of their estimates did not differ across languages. For example, although there was a significant difference in ordinality across French and English, this difference was very small (2%) and disappeared when analyses were restricted to numbers known in both French and English. This suggests that once children exhibit later-greater knowledge in one language, they show roughly equal competence in deploying it in their L2 for familiar numbers — a pattern that is compatible with the use of a language-general principle rather than language- and item-specific learning. In addition, we found no difference in the variability of estimates for familiar numbers across languages; although children mapped their number words to magnitudes less accurately in French than in English, their mis-mappings were nevertheless every bit as consistent.

One hypothesis that may explain this pattern of findings is structure mapping, a mechanism discussed in previous studies of number word learning (Carey, 2004, 2009; Gentner, 2010; Wynn, 1992), dot array estimation (Izard & Dehaene, 2008; Lyons, Ansari, & Beilock, 2012; Sullivan & Barner, 2013, 2014), and number line reasoning (Cohen & Sarnecka, 2014; Siegler, Thompson, & Schneider, 2011; Thompson & Opfer, 2010; for a review, see Marchand & Barner, 2018). On this view, mapping numerical symbols to magnitudes is not merely a process of forming item-based associations between individual words and ANS values but instead involves a process of noticing — and exploiting — the analogous structures of the count list and ANS values. For example, to make an accurate estimate for a dot array containing 20 items, a child who has a robust item-based association between 10 and the word *ten* might first use the later-greater principle to infer that because the 20-dot array is greater in number, it merits a label that comes later in the count list — thereby exploiting knowledge of the structure of counting (and a principle that might be transferred across languages, compatible with our finding that ordinality does not differ across children's

languages). For this estimate to be accurate, however, the child must further exploit the structure of counting in a way that is language specific. Using a process similar to anchoring and adjustment (Tversky & Kahneman, 1974), a child who knows (or guesses) that a particular dot array contains *ten* items and subsequently sees an array twice as large could make an estimate of *twenty*, provided that the child knows that *twenty* is twice as far into the count list as *ten*. Thus, the child must know not only the ordering of number words in his or her language but also the relative distance between number words in his or her count list — that is, knowledge that depends on language-specific mastery of the count list.

On this structure mapping hypothesis, there are multiple ways in which bilinguals might become better estimators in one language than in another language. First, they might acquire better knowledge of the relative distance between numbers in a particular language. Although the order of the count list makes the distance between numbers implicitly available to children, this knowledge is likely difficult to deploy, much as it is difficult for even adults to estimate the relative distance between letters of the alphabet (Klahr, Chase, & Lovelace, 1983). Knowledge that, for example, 40 is twice 20 likely requires experience beyond reciting the count list and may benefit from practice counting with decades or reciting basic multiplication facts. Second, differences in estimation across languages may be explained, in part, by differences in calibration. Although it is unlikely that all large number words have item-specific associative mappings to ANS values, a small number of anchor points for especially frequent numbers may exist (Sullivan & Barner, 2013, 2014). A child who lacks any anchors might know the later-greater principle and that 40 is twice 20 but might still systematically under- or overestimate due to mis-calibration (i.e., not having associative mappings to ANS values that act as anchors for larger estimates that rely on structural inference). Finally, also compatible with structure mapping, if bilingual children have

limited access to number words in their L2, they might stretch their response grid (Izard & Dehaene, 2008), thereby changing their estimates for all numbers that do not exhibit a strong associative mapping and causing the children to be less accurate for larger numbers even if they do not differ in ordinality or COV. Future studies should explore this question by directly probing children's structural knowledge of the count list and by testing the effects of calibration events on estimation accuracy across bilingual children's two languages.

Note that the ability to use the structure of counting to guide estimates may depend on how accessible this structure is in a particular language, which may be relevant to understanding differences between French and English estimation. In English, the words *seventy*, *eighty*, and *ninety* are relatively regular and are composed of the morphemes ''-ty" and ''seven," ''eight," and ''nine," respectively. In contrast, in French the word *soixante-dix* (70) translates as ''sixty-ten", *quatre-vingts* (80) translates as ''four-twenty", and *quatre-vingt-dix* (90) translates as ''four-twenty-ten". These structural irregularities may make these numbers harder to learn in French than in English, predicting poorer estimation. Conversely, they may make estimation easier by making the multiplicative relations between numbers more transparent. Unfortunately, in our dataset this question is difficult to test because we did not assess whether children had decomposed these words according to rules and also because the irregularities in French happen to arise specifically with larger numbers. Consequently, although estimates for these numbers could be poorer due to irregularities, less accurate estimates could also result from the fact that, in a weaker language larger numbers are simply not as familiar as they are in a child's primary language. Although no previous studies have examined the impact of counting regularity on estimation, past work has shown that children who are exposed to highly regular counting systems like Cantonese learn to count earlier than children learning less regular systems like Hindi and may also be quicker to

learn other numerical skills, although such effects are variable across studies and some attribute them to educational practices rather than language (Dowker, Bala, & Lloyd, 2008; Dowker & Roberts, 2015; Lefevre, Clarke, & Stringer, 2002; Mark & Dowker, 2015; Miller, Kelly, & Zhou, 2005; Miller, Smith, Zhu, & Zhang, 1995; Miller & Stigler, 1987; Miura, Kim, Chang, & Okamoto, 1988; Miura & Okamoto, 1989, 2013; Schneider et al., 2020).[3] However, these findings notwithstanding, there is reason to believe that irregularities in how languages represent decade terms such as *sixty* and *seventy* do not likely contribute to these differences. For example, although English- and Cantonese-speaking children readily learn rules to combine decade labels (10–90) with unit labels (1–9), children who speak these languages do not appear to know that decade labels are generated by rules and instead treat words like *sixty* as unanalyzed morphemes (as shown by the fact that, when asked to count, children frequently count up to decade transitions such as 29, 39, and 49 but cannot use a rule to generate decade labels; Schneider et al., 2020). Thus, evidence from other languages suggests that French-speaking children likely represent words like *soixante-dix* and *quatre-vingts* as unanalyzed words, such that their relative complexity, therefore, does not affect estimation either positively or negatively. Future studies should explore this question.

In summary, this study shows that estimation abilities differ across a bilingual child's two languages and that these differences are not explained by count list knowledge alone given that they persist even within numbers used in both languages. These results have potentially important consequences for mathematics education given that, although previous studies show associations

[3] Note that Lefevre et al. (2002) found that French Canadian children do not count as fluidly as English-speaking children but also that they do not receive as much counting input, making it unclear whether linguistic differences in the count list play a role.

between ANS acuity and mathematics achievement (Halberda, Mazzocco, & Feigenson, 2008; Libertus, Feigenson, & Halberda, 2011; Starr, Libertus, & Brannon, 2013), other studies find a link between math achievement and estimation abilities, with some studies arguing that this second relationship is stronger (Booth & Siegler, 2006, 2008; Gunderson, Ramirez, Beilock, & Levine, 2012; Kolkman, Kroesbergen, & Leseman, 2013; Moore & Ashcraft, 2015; Sasanguie, De Smedt, Defever, & Reynvoet, 2012; Siegler & Booth, 2004; Sullivan et al., 2016). These factors, combined with our findings, suggest that if caregivers, preschool teachers, and other educators wish to train estimation abilities as a mechanism for promoting later mathematics ability, it may be important that this training take place in the expected language of later mathematics instruction. Likewise, children who are likely to transition midway through their mathematics training from one modality of instruction (e.g., French) to another (e.g., English) may benefit from estimation training in both languages. In particular, bilingual children may benefit from being explicitly taught about language-specific aspects of estimation in their L2 such as the distance between numerals (e.g., counting by 10s, learning that 20 = double 10). Developing bilinguals' understanding of the structure of their counting systems in both languages may improve their estimation abilities and, more generally, their mappings between the linguistic and nonlinguistic representations of magnitudes.

Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in Marchand, E., Wade, S., Sullivan, J., and Barner D. (2020). Language-specific numerical estimation in bilingual children. *Journal of Experimental Child Psychology*, doi.org/10.1016/j.jecp.2020.104860. The dissertation author was the primary investigator and author of this paper.

This work received support from the Social Sciences and Humanities Research Council of Canada via a fellowship to E.M., a National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1419118 to S.W., and a James S. McDonnell Foundation award to D.B. Special thanks go to the participating children and families from Robb Road, Coeur-de-l'Ile, and La Petite École as well as research assistants Karen Bejar, Ariane Brunelle, Irene Chavez, Jessica Valdivia, Mikael Harvey, Rubí Hernández, and Miriam Rubenson. We also thank the members of the Language and Development Lab at UCSD and the reviewers for their helpful comments and feedback.

References

Agrillo, C., Dadda, M., Serena, G., & Bisazza, A. (2008). Do fish count? Spontaneous discrimination of quantity in female mosquitofish. *Animal cognition*, *11*(3), 495-503.

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, *86*(3), 201-221.

Barth, H., Starr, A., & Sullivan, J. (2009). Children's mappings of large number words to numerosities. *Cognitive Development*, *24*(3), 248-264.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014.

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental psychology*, *42*(1), 189.

Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. *Child development*, *79*(4), 1016-1031.

Chomsky, N. (2008). On phases. In R. Freidin, C. P. Otero and M. L. Zubizarreta (eds.) Foundational issues in linguistic theory. Essays in honor of Jean-Roger Vergnaud. Cambridge, MA: MIT Press. 134–166.

Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, *133*(1), 59-68.

Carey, S. (2009). Where our number concepts come from. *The Journal of philosophy*, *106*(4), 220.

Cohen, D. J., & Sarnecka, B. W. (2014). Children's number-line estimation shows development of measurement skills (not number representations). *Developmental psychology*, *50*(6), 1640.

Coppola, M., Spaepen, E., & Goldin-Meadow, S. (2013). Communicating about quantity without a language model: Number devices in homesign grammar. *Cognitive psychology*, *67*(1), 1-25.

Dagenbach, D., & McCloskey, M. (1992). The organization of arithmetic facts in memory: Evidence from a brain-damaged patient. *Brain and Cognition*, *20*(2), 345-366.

Davidson, K., Eng, K., & Barner, D. (2012). Does learning to count involve a semantic induction?. *Cognition*, *123*(1), 162-173.

Dehaene, S. (1997). The Number Sense. New York: Oxford University Press.

Dehaene, S., & Cohen, L. (1997). Cerebral pathways for calculation: Double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex*, *33*(2), 219-250.

Dehaene, S., Molko, N., Cohen, L., & Wilson, A. J. (2004). Arithmetic and the brain. *Current opinion in neurobiology*, *14*(2), 218-224.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, *284*(5416), 970-974.

Delazer, M., & Benke, T. (1997). Arithmetic facts without meaning. *Cortex*, *33*(4), 697-710.

DeSmedt, B., Noël, M. P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education*, *2*(2), 48-55.

Desoete, A., Ceulemans, A., De Weerdt, F., & Pieters, S. (2012). Can we predict mathematical learning disabilities from symbolic and non-symbolic comparison tasks in kindergarten? Findings from a longitudinal study. *British Journal of Educational Psychology*, *82*(1), 64-81.

Di Sciullo, A. M. (2012, February). Biolinguistics, minimalist grammars, and the emergence of complex numerals. In *Evolang IX. Workshop on Theoretical Linguistics/Biolinguistics* (pp. 13-18).

Dixon, R. M. W. (2004). *The Jarawara language of southern Amazonia.* Oxford: Oxford University Press.

Domahs, F., & Delazer, M. (2005). Some assumptions and facts about arithmetic facts. *Psychology Science, 47(1),*96-111.

Dowker, A., Bala, S., & Lloyd, D. (2008). Linguistic influences on mathematical development: How important is the transparency of the counting system?. *Philosophical Psychology, 21*(4), 523-538.

Dowker, A., & Roberts, M. (2015). Does the transparency of the counting system affect children's numerical abilities?. *Frontiers in psychology, 6,* 945.

Epps, P. (2006). Growing a numeral system: The historical development of numerals in an Amazonian language family. *Diachronica*, *23*(2), 259-288.

Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of experimental child psychology, 123*, 53-72.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*(7), 307-314.

Frank, M. C., & Barner, D. (2012). Representing exact number visually using mental abacus. *Journal of Experimental Psychology: General*, *141*(1), 134.

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition*, *108*(3), 819-824.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*(5), 752-775.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, *306*(5695), 496-499.

Gunderson, E. A., Ramirez, G., Beilock, S. L., & Levine, S. C. (2012). The relation between spatial skill and early number knowledge: the role of the linear number line. *Developmental psychology*, *48*(5), 1229.

Gunderson, E. A., Spaepen, E., & Levine, S. (2015). Approximate number word knowledge before the cardinal principle. *Journal of Experimental Child Psychology*, 130, 35–55.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the" Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental psychology*, *44*(5), 1457.

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665.

Hatano, G., Miyake, Y., & Binks, M. G. (1977). Performance of expert abacus operators. *Cognition*, *5*(1), 47-55.

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of experimental child psychology*, *103*(1), 17-29.

Ifrah, G. (2000). The universal history of numbers: From prehistory to the invention of the computer, translated by David Vellos, EF Harding, Sophie Wood and Ian Monk.

Ischebeck, A., Zamarian, L., Siedentopf, C., Koppelstätter, F., Benke, T., Felber, S., & Delazer, M. (2006). How specifically do we learn? Imaging the learning of multiplication and subtraction. *Neuroimage*, *30*(4), 1365-1375.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221-1247.

Klahr, D., Chase, W. G., & Lovelace, E. A. (1983). Structure and process in alphabetic retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(3), 462.

Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. (2013). Early numerical development and the role of non-symbolic and symbolic skills. *Learning and instruction*, *25*, 95-103.

Lampl, Y., Eshel, Y., Gilad, R., & Sarova-Pinhas, I. (1994). Selective acalculia with sparing of the subtraction process in a patient with left parietotemporal hemorrhage. *Neurology*, *44*(9), 1759-1759.

Le Corre, M. (2014). Children acquire the later-greater principle after the cardinal principle. *British Journal of Developmental Psychology, 32*(2), 163-177.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: an investigation of the conceptual sources of the verbal counting principles. *Cognition*, 105(2), 395–438.

Lee, K. M. (2000). Cortical areas differentially involved in multiplication and subtraction: a functional magnetic resonance imaging study and correlation with a case of selective acalculia. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *48*(4), 657-661.

Lefevre, J. A., Clarke, T., & Stringer, A. P. (2002). Influences of language and parental involvement on the development of counting skills: Comparisons of French-and English- speaking Canadian children. *Early Child Development and Care, 172*(3), 283-300.

Lemer, C., Dehaene, S., Spelke, E., & Cohen, L. (2003). Approximate quantities and exact number words: Dissociable systems. *Neuropsychologia*, *41*(14), 1942-1958.

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental science*, *14*(6), 1292- 1300.

Lipton, J. S., & Spelke, E. S. (2005). Preschool children's mapping of number words to nonsymbolic numerosities. *Child Development*, *76*(5), 978-988.

Lyons, I. M., Ansari, D., & Beilock, S. L. (2012). Symbolic estrangement: Evidence against a strong association between numerical symbols and the quantities they represent. *Journal of Experimental Psychology: General*, *141*(4), 635.

Marchand, E., & Barner, D. (2018). Analogical Mapping in Numerical Development. In *Language and Culture in Mathematical Cognition* (pp. 31-47). Academic Press.

Mark, W., & Dowker, A. (2015). Linguistic influence on mathematical development is specific rather than pervasive: revisiting the Chinese Number Advantage in Chinese and English children. *Frontiers in Psychology, 6*, 203.

Mejias, S., & Schiltz, C. (2013). Estimation abilities of large numerosities in Kindergartners. *Frontiers in psychology*, *4*.

Menninger, K. (1969). Number words and number symbols: A cultural history of numbers. Courier Corporation.

Miller, K. F., Smith, C. M., Zhu, J., & Zhang, H. (1995). Preschool origins of cross-national differences in mathematical competence: The role of number-naming systems. *Psychological Science, 6*(1), 56-60.

Miller, K. F., & Stigler, J. W. (1987). Counting in Chinese: Cultural variation in a basic cognitive skill. *Cognitive Development*, 2(3), 279-305.

Miller, K. F., Kelly, M., & Zhou, X. (2005). Learning mathematics in China and the United States: Cross-cultural insights into the nature and course of preschool mathematical development. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 163- 177). New York, NY, US: Psychology Press.

Miura, I. T., Kim, C. C., Chang, C. M., & Okamoto, Y. (1988). Effects of language characteristics on children's cognitive representation of number: Cross-national comparisons. *Child Development,* 1445-1450.

Miura, I. T., & Okamoto, Y. (1989). Comparisons of U.S. and Japanese First Graders' Cognitive Representation of Number and Understanding of Place Value. *Journal of Educational Psychology, 81*(1), 109–114.

Miura, I. T., & Okamoto, Y. (2013). Language supports for mathematics understanding and performance. In *The development of arithmetic concepts and skills* (pp. 251-264). Routledge.

Moore, A. M., & Ashcraft, M. H. (2015). Children's mathematical performance: Five cognitive tasks across five grades. *Journal of experimental child psychology*, *135*, 1-24.

Platt, J. R., & Johnson, D. M. (1971). Localization of position within a homogeneous behavior chain: Effects of error contingencies. *Learning and Motivation*, *2*(4), 386-414

Pesenti, M., Seron, X., & Van Der Linden, M. (1994). Selective impairment as evidence for mental organisation of arithmetical facts: BB, a case of preserved subtraction?. *Cortex*, *30*(4), 661-671.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, *306*(5695), 499-503.

Saalbach, H., Eckstein, D., Andri, N., Hobi, R., & Grabner, R. H. (2013). When language of instruction and language of application differ: Cognitive costs of bilingual mathematics learning. *Learning and Instruction*, *26*, 36-44.

Sasanguie, D., De Smedt, B., Defever, E., & Reynvoet, B. (2012). Association between basic numerical abilities and mathematics achievement. *British Journal of Developmental Psychology*, *30*(2), 344-357.

Scarf, D., Hayne, H., & Colombo, M. (2011). Pigeons on par with primates in numerical competence. *Science*, *334*(6063), 1664-1664.

Schneider, R. M., Sullivan, J., Marušič, F., Biswas, P., Mišmaš, P., Plesničar, V., & Barner, D. (2020). Do children use language structure to discover the recursive rules of counting?. *Cognitive Psychology,* 117, 101263.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child development*, *75*(2), 428-444.

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive psychology, 62*(4), 273-296.

Spaepen, E., Coppola, M., Spelke, E. S., Carey, S. E., & Goldin-Meadow, S. (2011). Number without a language model. *Proceedings of the National Academy of Sciences*, *108*(8), 3163- 3168.

Spaepen, E., Coppola, M., Flaherty, M., Spelke, E., & Goldin-Meadow, S. (2013). Generating a lexicon without a language model: Do words for number count?. *Journal of memory and language*, *69*(4), 496-505.

Spelke, E. S. (2017). Core knowledge, language, and number. *Language Learning and Development*, *13*(2), 147-170.

Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition*, *78*(1), 45-88.

Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences*, 110(45), 18116-18120.

Sullivan, J., & Barner, D. (2013). How are number words mapped to approximate magnitudes?. *The Quarterly Journal of Experimental Psychology*, *66*(2), 389-402.

Sullivan, J., & Barner, D. (2014). Inference and association in children's early numerical estimation. *Child development*, *85*(4), 1740-1755.

Sullivan, J., Frank, M. C., & Barner, D. (2016). Intensive math training does not affect approximate number acuity: Evidence from a three-year longitudinal curriculum intervention. *Journal of Numerical Cognition*, *2*(2), 57-76.

Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development*, *81*(6), 1768-1786.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1130.

van Harskamp, N. J., & Cipolotti, L. (2001). Selective impairments for addition, subtraction and multiplication. Implications for the organisation of arithmetical facts. *Cortex*, *37*(3), 363-388.

van Harskamp, N. J., Rudge, P., & Cipolotti, L. (2002). Are multiplication facts implemented by the left supramarginal and angular gyri?. *Neuropsychologia*, *40*(11), 1786-1793.

Wagner, K., Kimura, K., Cheung, P., & Barner, D. (2015). Why is number word learning hard? Evidence from bilingual learners. *Cognitive psychology*, *83*, 1-21.

Wagner, J. B., & Johnson, S. C. (2011). An association between understanding cardinality and analog magnitude representations in preschoolers. *Cognition*, *119*(1), 10-22.

Watanabe, A. (2017). Natural language and set-theoretic conception of natural number. *Acta Linguistica Academica*, *64*(1), 125-151.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155-193.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive psychology*, *24*(2), 220-251.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1-B11.

Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental science*, *8*(1), 88-101.

CHAPTER 2

Assessing the knower-level framework: How reliable is the Give-a-Number task?

Elisabeth Marchand, Jarrett Lovelett, Kelly Kendro and David Barner[1]

[1]Department of Psychology, University of California, San Diego

Abstract

The Give-a-Number task has become a gold standard of children's number word comprehension in developmental psychology. Recently, researchers have begun to use the task as a predictor of other developmental milestones. This raises the question of how reliable the task is, since test-retest reliability of any measure places an upper bound on the size of reliable correlations that can be found between it and other measures. In Experiment 1, we presented 81 2- to 5-year-old children with Wynn's (1992) titrated version of the Give-a-Number task twice within a single session. We found that the reliability of this version of the task was high overall, but varied importantly across different assigned knower levels, and was very low for some knower levels. In Experiment 2, we assessed the test-retest reliability of the non-titrated version of the Give-a-Number task with another group of 81 children and found a similar pattern of results. Finally, in Experiment 3, we asked whether the two versions of Give-a-Number generated different knower levels within-subjects, by testing 75 children with both tasks. Also, we asked how both tasks relate to another commonly used test of number knowledge, the "What's-On-This-Card" task. We found that overall, the titrated and non-titrated versions of Give-a-Number yielded similar knower levels, though the non-titrated version was slightly more conservative than the titrated version, which produced modestly higher knower levels. Neither was more closely related to "What's-On-This-Card" than the other. We discuss the theoretical and practical implications of these results.

Introduction

Over the past 40 years, a large corpus of studies has shown that children acquire the meanings of number words in a predictable and protracted stage-like sequence. This developmental sequence has been revealed in large part by a single measure of number word knowledge, called the Give-a-Number task (Give-N). Though versions of this task were used as early as the 1970s to study number word comprehension (Schaeffer et al., 1974), Give-N emerged as a type of gold standard after it was used by Wynn (1990, 1992) to describe children's progression through stage-like "knower levels" in both cross-sectional and longitudinal designs. In the task, an experimenter provides children with a set of small counters (e.g., 10-15 toy apples), and asks them to give specific numbers of things, often starting with 1 – e.g., "Can you put *one* apple in the plate?". Children who can consistently give 1 when asked for *one*, but who give inconsistent amounts of objects for other requests are typically called 1-knowers. Similarly, 2-knowers can give 1 and 2 when asked for these quantities but are unable to consistently give appropriate quantities for larger numbers like *three, four,* etc. Following a similar pattern, children go through the stages of 3-knower and sometimes 4-knowers, too. Sometime between the ages of 3;6 and 5, children appear to make a breakthrough, and begin to use counting to correctly give larger sets, at which point they are called "Cardinal Principle knowers" or CP-knowers. This basic developmental pattern appears to be highly replicable across multiple labs in different countries (Almoammer et al., 2013; Barner, Libenson & Yang, 2009; Ceylan & Aslan, 2018; Condry & Spelke, 2008; Davidson et al., 2012; Jara-Ettinger et al., 2017; Le Corre & Carey, 2007; Le Corre et al., 2006, 2016; Li et al., 2003; Marchand & Barner, 2019; Meyer et al., 2020; Negen & Sarnecka, 2012; Nikoloska, 2009; Piantadosi et al., 2014; Sarnecka & Carey, 2008; Sarnecka et al., 2007; Sarnecka & Lee, 2009; Sarnecka, et al., 2018; Schneider et al., 2020; Slusser et al., 2013;

Spaepen et al., 2018; Wagner et al., 2015; Wynn, 1990, 1992). This is important not only because of the theoretical implications of the observed stages (e.g., Carey & Barner, 2019; Le Corre & Carey, 2007; Le Corre et al., 2006; Piantadosi et al., 2012; Sarnecka & Carey, 2008; Sella et al., 2020), but also because the stages provide a framework for comparing data across studies and across cultures. Numerous studies have now tested how vocabulary size, grammatical cues, and other cultural factors relate to different knower level stages (Almoammer et al., 2013; Barner et al., 2009; Le Corre et al., 2016; Marušič et al., 2016; Negen & Sarnecka, 2012; Sarnecka et al., 2007, 2018), and others have asked how knower levels relate to later mathematics achievement (Chu et al., 2016; Geary & vanMarle, 2016; Moore et al., 2016; Purpura & Simms, 2018; Spaepen et al., 2018) or the development of other cognitive processes (Abreu-Mendoza et al., 2013; Le Corre, 2014; Mussolin et al., 2014; Sarnecka & Wright, 2013; Shusterman et al., 2016).

Critically, however, the replicability of the overall knower level framework does not itself assure the reliability of individual knower level classifications and doesn't guarantee that testing correlations between knower levels and other factors will generate meaningful results. Currently, the reliability of the Give-N task is not known. This is important because the strength of a correlation between two observations (e.g., knower level and vocabulary size), *r*(ObservedA,ObservedB), is bounded not only by the true correlation between the true value of the variables being measured, *r*(TrueA,TrueB), but also by the test-retest reliability of these measures taken individually, reliabilityA, reliabilityB (Nunnally, 1970).

$$r(ObservedA, ObservedB = r(TrueA, TrueB) \times \sqrt{(reliabilityA \times reliabilityB)}$$

Thus, as noted by Vul et al. (2009), in a scenario in which a true correlation between two variables is 100% but the test-retest reliability is .7 for one measure and .8 for the second, the highest detectable correlation should be .75 (i.e., 1 x √(.7 x .8)). In the current context, this means

that if individual knower levels (e.g., the 1-knower stage) exhibit very low reliability (e.g., .3), then the size of expected correlations between this knower level and other variables should also be low. Consequently, very low reliability would draw into question the validity of knower levels, since the validity of a measure is defined by its ability to make predictions about the outcomes of other measures.[4] More generally, the interpretation of knower level assignments as correlates of other outcomes hinges critically on the reliability of the Give-N task.

In the present study, we investigated the reliability of the Give-N task in three experiments. In Experiment 1, we assessed the test-retest reliability of Wynn's titrated version of Give-N. In the titrated Give-N task, trials are structured such that if a child responds correctly to a request (e.g., giving exactly 1 object when asked for *one*), they are then tested with the next largest number (e.g., *two*), whereas if they fail, they are tested on a smaller number (or again on *one*). This procedure is then repeated until the experimenter can identify the largest number known by the child. In Experiment 2, we investigated the test-retest reliability of an alternative version of Give-N that uses a non-titrated trial structure in which children are tested on all numbers of interest (e.g., 1, 2, 3, 4, 5, 6, 8, 10) three times each in pseudo-random order, using the same criteria to identify children's knower levels. We expected that this version might offer stronger reliability than the titrated version, because it features more trials and uses the same trial structure on each testing occasion, unlike the titrated version.[5] In both Experiments 1 and 2, we also considered the role that

[4] As explained in Buelow (2020): "A task that is not reliable can not be valid, and lowered reliability can limit inferences made from the task to real-world behaviors." The logic is that an outcome X can't predict a second outcome Y if it can't predict itself (i.e., if it is unreliable). And if X can not explain properties of the world, then it is not a valid measure.

[5] We reasoned that additional trials might increase reliability by providing more information and reducing the likelihood of underestimating (or overestimating) knowledge, and that stable trial structure should reduce the possibility that low reliability is due to variability introduced by differences in methods across testing sessions. Specifically, whereas random performance errors made by children will have no impact on the trial structure of the non-titrated version of the task since its trial structure is predetermined, such errors may significantly change the trial structure of the titrated version, since an error on a trial forces a retreat to a smaller number.

testing environment might play in the reliability of Give-N by evaluating children in two different settings – either in the lab or outside of the lab (e.g., in a museum, preschool, etc.) – as some studies have reported different outcomes in these different settings (Newman et al., 1978; Rasmussen et al., 2017; Yantz & McCaffery, 2009; cf. Pfefferle et al., 1982). Finally, in Experiment 3, we compared the titrated and non-titrated Give-N versions within-subjects, to determine whether they generated different results, and whether either of the two was more conservative (e.g., by ascribing less knowledge). Also, Experiment 3 attempted to probe how the two Give-N methods are related to another frequently used measure of number knowledge by comparing them to the What's-on-this-Card task, which assesses how accurately children label sets when presented visually.

Experiment 1: Give-a-Number Titrated

Method

Participants

We tested 106 English-speaking children. A total of 25 children were excluded from analysis because of (1) failure to complete all 3 tasks (n = 11), (2) language delay (n = 1), (3) being a non-English primary speaker (n = 2), (4) falling outside the targeted age range (n = 4) or (5) experimenter error (n = 7). Our final sample included 81 children, aged 2;2 to 4;1-year-old (M = 3;3 years). We chose to test participants in this age range as previous studies suggest that it features the most variability in knower levels. Participants were recruited from a parent database (lab), preschools, and museums in San Diego, California, spanning a wide range of socioeconomic backgrounds. Informed consent was obtained from parents. The study received approval by the institutional ethics committee of UCSD.

Materials and procedure

Children were tested either in the lab or offsite at museums and preschools. The testing environment in museums and preschools was similar and consisted of a relatively quiet corner of a room made available by staff. The testing environment in the lab was more quiet than off-site and possible distractions were limited. Each session lasted approximately 8 minutes and included three tasks administered in the same order for all participants: (1) Give-a-Number task 1, (2) Highest Count task and (3) Give-a-Number task 2. Children received a small prize for their participation at the end of the testing session.

Titrated Give-a-Number Task

This task was based on Wynn (1992). Stimuli included a puppet, a plastic plate, and a pile of small plastic toys. Participants were asked to provide a certain number of toys in the following way: "*Mr. Monkey is very hungry. This is a plate and these are your bananas. I want you to put bananas on the plate for Mr. Monkey, ok? Listen carefully! Can you put N banana(s) on the plate?* (N is the number word). *Put N banana(s) on the plate and tell me when you're all done.*" Following these instructions, children were asked to count to verify that they had provided N (i.e.,*"Is that N? Can you count and make sure?"*). If they chose to change their answers, only their final responses were recorded. Participants were always asked for *one* first, and then *two*. If the child succeeded on both trials, the experimenter then asked for *three.* Otherwise, they asked for *one*. The subsequent requests depended on the child's pattern of response: if the child succeeded in providing N items, the experimenter asked for N + 1 and if the child failed, they asked for N - 1. The lowest request was *one* and the highest was *six*. Children were credited as N-knowers (e.g., 2-knowers) if they correctly gave N objects at least 67% of the time when asked for N. Furthermore, to be credited as N-knowers, children needed to use N 67% of the time only for requests of N and not for other requests (in practice, this meant that children could give N only once for requests

other than N). Children were credited as CP-knowers if they were able to provide all sets up to *six* based on these criteria, or if they responded to each request (*one* to *six*) consecutively without error, in accordance with Sarnecka and Wright (2013). Aside from this last instance (of CP-knowers), participants were tested with a minimum of 2 trials for *N,* and for numbers tested twice, children needed to succeed on both trials to be tested on the next trial or be credited as N-knower. Children who correctly gave 1 object when asked for *one* (but failed for *two* and larger requests) were classified as 1-knowers. Children who answered successfully for *one* and *two* were credited as 2-knowers and so forth. Although past studies have often classified children who succeed at *five* as CP-knowers, we chose to categorize children as 5-knowers if they succeeded at *five* but failed at *six*. Although more conservative, this criterion allowed us to test the claim that knower levels higher than 4 exist and can be diagnosed (Krajcsi et al., 2018). However, allowing for an additional knower level in the classification risks decreasing the reliability of the task and of some knower levels in particular. Nevertheless, as we report below, including 5-knowers didn't impact the reliability of the task because the number of 5-knowers was low (3 children at T1 and 4 at T2).[6]

Highest Count (HC)

This task was used to verify that our sample was representative of previously reported samples of the same age and served as a filler task between the two Give-N tests. Participants were asked to count as high as they could. The last number reached before stopping or making an error was recorded as the child's highest count.

Analyses

[6] We re-ran all analyses with 5-knowers categorized as CP-knowers and obtained virtually the same results as presented below; while the reliability for the task overall remained unchanged (linear weight = 0.88 vs 0.87), there was, unsurprisingly, a slight increase of reliability for the CP-knowers (from 0.827 to 0.852), the subset-knowers (from 0.681 to 0.700) and the knower-level groups (0.824 to 0.850) analyses. However, in all of these cases, the increase was negligible.

The choice of a reliability index depends crucially on the scale of the outcome measure of interest. Cohen's Kappa is very commonly used for nominal scales, especially when the outcome of interest is binary, such as the presence or absence of some clinical condition (Hallgren, 2012). However, the basic computation of Kappa can be modified to weight different disagreements in classification differently, allowing the approach to work for ordinal scales as well (in which, say, the difference between 4 and 2 is larger than that between 4 and 3). Intra-class correlations (ICC; Hallgren, 2012) are designed for use with tasks that produce continuous outcome measures, but also produce interpretable results for ordinal scales. Thus to select a measure of reliability for knower level classifications, one must first decide how to conceptualize that construct: as a smooth continuum of knower levels, or as a discrete set of stages? Is the transition from zero-knower to one-knower a similar jump in number knowledge as the transition from two-knower to three-knower? Because it is not clear whether any single choice of reliability metric is entirely free of drawbacks with respect to complexity of the knower level scale, we introduce and report several different metrics, so that readers may use their own judgment in assessing the degrees of reliability reported here.

Here we describe the different reliability indexes used throughout Experiments 1, 2 and 3, including Kappa (weighted and unweighted), Agreement, Bias Index, Prevalence Index, Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) and Intra-class correlation (ICC). The Kappa statistic was preferred for our reliability assessment as it is considered a standardized index of reliability for categorical variables (Hallgren, 2012), with which the knower level framework is more compatible (relative to continuous scales). Across the different measures of reliability for categorical data, we also prioritized the Kappa statistic because Kappa, in its weighted version, is compatible with both nominal and ordinal data, which we used in our experiments, allowing us

therefore to provide a consistent measure across different analyses. However, in addition to the weighted Kappa, we provide the reader with Intra-class Correlation when dealing with ordinal data as it is a common measure used in this context and may be preferred by researchers who conceptualize knower levels as a continuous scale (although we do not endorse this practice). All analyses were computed in R (Team R Code, 2018) and Kappa analyses were performed using the "vcd" (Meyer et al., 2021) and epiR packages (Stevenson & Sergeant, 2021). In our main analyses, reliability was measured using the weighted version of the Kappa statistic (Cohen, 1960; 1968), defined in the following way:

$$K = \frac{P_o - P_e}{1 - P_e}$$

In this expression, K represents the Kappa statistic, Po is the observed (overall) agreement and Pe the agreement expected by chance. Overall agreement corresponds to the total number of matches between the first and second assessment of a task (i.e., the sum of the values on the diagonal of a contingency table) divided by the total number of observations (see Table 2.1 for an example of a simplified contingency table). Agreement expected by chance refers to the sum of the theoretical frequencies in each cell of the diagonal, which are calculated using the same formula as for computing expected frequencies for Pearson's Chi square (i.e., by taking the product of observed marginal proportions classified as each knower level across tasks).

In the modified weighted Kappa formula, Po and Pe are calculated using a matrix of (dis-)agreement weights, which specify the degree to which each possible pair of classifications from the two tasks (dis-)agree. In the case of knower levels, this means that the difference between, for example, a 1-knower and a 5-knower can be represented as larger than the difference between a 1-knower and a 2-knower. That feature enables weighted Kappa to handle ordinal scales, since it can attach greater weight to large differences between levels than to small differences (Cohen, 1968).

Importantly, it is incumbent on the investigator to decide *how much* weight to assign each possible (dis-)agreement, by carefully designing a weight matrix.

In principle, a fully custom weight system could be used to describe the severity of disagreement for each pairwise combination of classifications across the two tasks. For example, disagreements in which a subject is classified once as a CP-knower and once as a non-knower (CP-0k disagreements) could be weighted as arbitrarily more severe than disagreements in which the particular value of subset-knower was different across tasks. Each other combination of disagreements, such as CP-3k, 0k-4k, 4k-0k, etc., would have to be specified individually. *Any* choice of weighting, especially a custom weight scheme, therefore reflects a judgement regarding the nature of the number word acquisition process and its stages. For that reason, we refrain from developing our own customized weight system (with which other researchers could reasonably disagree). Instead, we report results using two common weighting systems: linear weights (in which the penalty for a disagreement is proportional to the absolute value of the difference in ranks across the two levels), and quadratic weights (in which the penalty is proportional to the square of that difference); using linear weights, a 4k-2k disagreement is twice as severe as a 4k-3k disagreement, while under quadratic weighting, 4k-2k is four times as severe as 4k-3k. Our preference is for linear weights, as that approach makes fewer theoretical assumptions about the trajectory of number knowledge development.

In addition to Kappa, in all analyses we reported either the overall agreement or the effective agreement depending on the data under study. Effective agreement is defined as the number of matches divided by the number of observations that include at least one of the knower levels in consideration. Both overall agreement, the total number of matches over total values, and effective agreement are inflated indexes of reliability because they don't consider the agreement

that could have occurred by chance (Luck et al., 2012; Viera & Garrett, 2005), which Kappa (weighted and unweighted) accounts for, making Kappa more conservative than raw measures of agreement.

Some authors have argued that the magnitude of Kappa can be influenced by factors such as prevalence and bias in the data and that consequently, Kappa can be misleading in cases where these factors are considerable (Byrt et al., 1993; Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990). Prevalence refers to the relative difference of agreement between raters or tasks across conditions. The prevalence index is calculated in the following way (refer to Table 2.1):

$$Prevalence\ index = \frac{|a - d|}{n}$$

Where $|a - d|$ is the absolute value of the difference between the frequencies of cells on the diagonal (agreements; *a* and *d* in the Table 2.1) and n is the total number of observations.

Table 2.1: Example of a simplified contingency table used in the reliability computations

		Assessment 1	
		1-knower	2-knower
Assessment 2	1-knower	a	c
	2-knower	b	d

Note. Example of contingency table with Give-N's One-Knower (1-knower) and Two-knower (2-knower) only.

If the prevalence index is high, suggesting that there is a high asymmetry in the frequencies in the cells of the diagonal, then Kappa will be reduced. The bias effect on Kappa occurs when there is large asymmetry in the frequencies of cells outside the diagonal, in other words, of disagreements (*b* and *c*). A high bias index can lead to an oversized Kappa. The bias index is measured in the following way:

$$Bias\ index = \frac{|b - c|}{n}$$

In our assessment of reliability, alongside Kappa, we provide the prevalence and bias indexes[7], as well as the Prevalence-adjusted bias-adjusted Kappa (PABAK) coefficient whenever the data allow it. It is important to note, however, that the prevalence and bias indexes are not measures of reliability per se, but instead provide an indication of potentially unbalanced data, and consequently, whether to rely more on PABAK than Kappa when interpreting the results.[8] PABAK is an adjustment of Kappa that takes into account the influence of bias and prevalence. It is calculated by substituting the actual frequencies of cells *a* and *d* by their average to account for prevalence, and by replacing the actual frequencies of cells *c* and *b* by their average to account for bias. Not all studies agree on which Kappa coefficient, the original or the PABAK, should be used as the main reference value. Some argue that bias and prevalence are the inevitable result of the natural disparities in the population under study and that correction coefficients such as PABAK can therefore be misleading. We follow the recommendations of Byrt and colleagues (1993) and provide the reader with both values (non-adjusted and PABAK) as well as the prevalence and bias indexes, whenever the data allowed, so that the reader can assess reliability based on a holistic evaluation of these measures. Finally, for some analyses when it was applicable, we also provided the ICC which is another commonly-used statistic for ordinal variables (Hallgren, 2012), based on

[7] Note that for tables larger than 2 × 2, we calculated the prevalence index by taking the average difference (in absolute value) between all numbers in the diagonal paired together (a-d in Table 2.1). More precisely, we replaced the |a-d| in prevalence formula by the average difference between all numbers on the diagonal of the contingency table under study. For the bias index, we replaced b in the formula by the sum of all numbers above the diagonale and c by the sum of all numbers below the diagonale. However, the literature on how to calculate these measures for tables larger than 2 × 2 was very sparse and we could not identify any straightforward way to proceed. Calculating the average prevalence and bias seemed like the more reasonable approach but other researchers might disagree. Results for the prevalence and bias indexes, as well as PABAK (which relies on these 2 measures) for large tables (>2 × 2) should therefore be interpreted with caution.

[8] The criteria to classify a prevalence and bias index as too high are subjective and inconsistent. In our data, we noticed that the prevalence index tended to be particularly high in 2 × 2 tables (e.g., 0.78) and in those cases, we favored the PABAK instead of Kappa for our interpretation of the data.

correlations. Our complete datasets are also available in the following repository: https://osf.io/48mke/.

Results

Table 2.2 shows the distribution of knower levels in the first and second assessment of the titrated Give-N task. On average children could count just above 10 in the Highest Count task (*M* = 12.8), and their counting skills were variable (*range* = 0 to 100; *SD* = 13.0). Seventy-four out of 81 (91%) participants were found to have a highest count greater than their knower level across the two Give-N assessments.

In our first analysis, we included all knower levels (non-knower to CP-knower) in a 7x7 contingency table (see Figure 2.1) and obtained an overall agreement of 77% and a weighted Kappa ($K_{w\text{-}linear}$) of .87 and .95 ($K_{w\text{-}quadratic}$; $Kappa_{unweighted}$ = .71; $Prevalence\ index_{(mean)}$ = .11, *range* = 0 - .25; *Bias index* = .09; $PABAK_{weighted}$ = .73; *ICC* = .97). All statistics are summarized in Table 2.3. Some researchers attempt to classify reliability scores according to a scale as described in Table 2.4; according to this scheme, this level of reliability is considered almost perfect (Landis & Koch, 1977; Fleiss et al., 2003). However, because there is disagreement regarding these labels and their utility (e.g., Sim & Wright, 2005), and because we are mainly interested in quantitative impacts of reliability on the size of correlations between measures (rather than qualitative endorsement of particular tasks), we sidestep the significance of these labels in our discussion. As shown in Figure 2.1, the rate of effective agreement (in percentage) across different knower levels was highly variable. Effective agreement was relatively high for non-knowers (80%), CP-knowers (76%), 1-knowers (71%), and 2-knowers (72%), but much lower for 3- (30%), 4- (18%) and 5-knowers (40%). Thus, overall, the titrated Give-N task was highly reliable when all knower levels were

considered together, although concordance between individual knower levels was lower in some cases.

To further investigate the difference of reliability across individual knower levels, we conducted three follow-up analyses for the subset-knower, non-knower, and CP-knower groups respectively. For the subset-knower analysis, we created a 6x6 contingency table with the knower levels 1 to 5, as well as a new category of non-subset-knowers (binning together non-knowers and CP-knowers) for Give-N Test 1 (T1) and Give-N Test 2 (T2). We found an effective agreement of 63% and an unweighted Kappa of .68 (*Prevalence index*$_{(mean)}$ = .15, *range* = 0 - .35; *PABAK* = .72).[9] We report the effective agreement here as we were interested in the agreement specifically within the group of subset-knowers and this index does not include non-knowers and CP-knowers. We also report the unweighted Kappa, and not the weighted Kappa, because weighted Kappa assumes an ordered category structure, which is violated by binning non-knowers and CP-knowers into a common category. Next, for the non-knower analysis, we generated a 2x2 contingency table with contrasting non-knowers with all other levels for both Give-N T1 and Give-N T2. We obtained an effective agreement of 80% and a Kappa of .88 and (*Prevalence index* = .78;[10] *Bias index* = 0; *PABAK* = .95). Next, for the CP-knower analysis, we created a 2x2 table (CP vs non-CP at T1 and T2) and found an effective agreement of 76% and a Kappa of .80 (*Prevalence index* = .37; *Bias index* = .04; *PABAK* = .83). These results suggest that the non-knower and CP-knower classifications are highly reliable, and more reliable than classifications within the subset stage

[9] Note here that the Bias Index is not valid in this subset-knowers analysis since the "non-subset-knower" category is not ordered and we would obtain different indexes based on its position in the contingency table, which can be placed arbitrarily either on the right or left of the contingency table.

[10] Note that in this analysis with non-knowers, the prevalence index is notably high and that using the PABAK coefficient as the main measure of reliability is recommended.

(though as already noted, concordance within the subset stage varies between individual levels, as shown in Figure 2.1).

In some past studies (e.g., Sarnecka & Carey, 2008), researchers have been less interested in whether a child is a specific N-knower (e.g., 1-knower), and more interested in whether they are CP-knowers or subset-knowers. Relatedly, most studies simply lack the power to analyze individual knower levels as predictors. In our next analyses, we therefore asked whether a child classified as, for example, a subset-knower in T1, was likely to be a subset-knower again in T2. To do this, we divided knower levels into three groups: non-knowers, subset-knowers (1K to 5K) and CP-knowers. We then created a 3x3 contingency table with knower level groups at T1 and knower level groups at T2. Here, we found an overall agreement of 89%, and a weighted Kappa (linear) of .82 and .86 (quadratic; $Kappa_{unweighted}$ = .80; $Prevalence\ index_{(mean)}$ = .28, *range* = .17 - .42; *Bias index* = .04; $PABAK_{weighted}$ = .83; *ICC* = .92), which is similar to the reliability of all knower levels taken together. This suggests that children who were classified as subset-knowers in the first assessment were very likely to remain subset-knowers in the second assessment, just like non-knowers and CP-knowers.

Next, we asked whether knower levels systematically increased or decreased between T1 and T2. An increase could signal a practice effect while a decrease could suggest a fatigue effect. In total, more children exhibited a decrease in their knower level from T1 to T2 (decreased n = 13; increased n = 6) but this difference was not significant (Wilcoxon rank test; W = 3368.5; p = .76). Furthermore, most of these children had knower levels that differed by one level (difference of 1 level, n = 11; difference of 2, n = 8).

Table 2.2: Distribution of Knower Levels at the First (T1) and second (T2) assessment of titrated Give-N

Knower Levels	0K	1K	2K	3K	4K	5K	CP
Assessment	Number of Participants						
Time 1	9	14	16	5	7	3	27
Time 2	9	15	15	8	6	4	24

Note. 0K refers to non-knower, 1K to 1-knower, 2K to 2-knower, 3K to 3-knower, etc, and CP to Cardinal Principle knower. In task 1, there were 9 children classified as non-knowers, 45 subset-knowers (1K to 5K) and 27 CP-knowers. In task 2, there were 9 non-knowers, 48 subset-knowers and 24 CP-knowers.

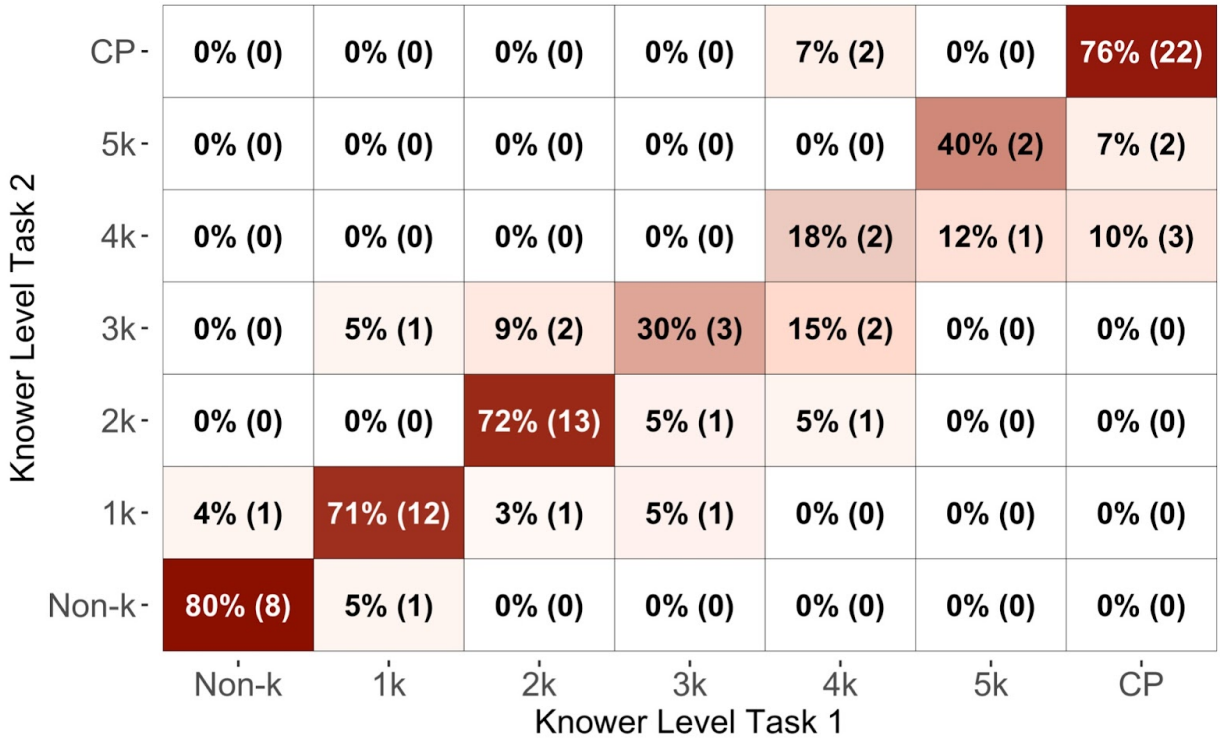

Figure 2.1: Knower Level Classification in the First and Second Assessments of Titrated Give-N

Note. The first assessment (T1) appears on the x axis, and the second assessment (T2) appears on the y axis. The percentages represent the percent effective agreement – i.e., the agreement calculated over not all paired knower levels, but those paired knower levels in which at least one belongs to the knower level in consideration. The numbers in parentheses represent the frequency of the paired knower level. The color scale is based on the proportion of effective agreement, where darker red represents higher agreement.

Table 2.3: Summary of Reliability measures and coefficients of the Titrated Give-N at T1 and T2 across different knower levels analyses

Group	Contingency table size	Agreement	K	PI	BI	PABAK	ICC
All knower levels	7 x 7	77%	.87 *(w-l) 95% CI,* .81 to .93 .95 *(w-q) 95% CI,* .92 to .98 .71 *(unw) 95% CI,* .59 to .82	.11 *(M) range:* 0 - .25	.09	.73 *(w-l) 95% CI,* .60 to .85 .73 *(w-q) 95% CI,* .56 to .89	.97 *95% CI,* .96 to .98
Subset-knower only	6 x 6	63% (effective)	.68 *(unw) 95% CI,* .56 to .80	.15 *(M) range:* 0 - .35	*NA*	.72 *(unw) 95% CI,* .61 to .83	*NA*
Non-knower vs others	2 x 2	80% (effective)	.88 *(unw) 95% CI,* .66 to 1	.78	.0	.95 *(unw) 95% CI,* .83 to .99	*NA*
CP-knower vs others	2 x 2	76% (effective)	.80 *(unw) 95% CI,* .58 to 1	.37	.04	.83 *(unw) 95% CI,* .66 to .93	*NA*
Knower-level groups	3 x 3	89%	.82 *(w-l) 95% CI,* .71 to .94 .86 *(w-q) 95% CI,* .76 to .95 .80 *(unw) 95% CI,* .68 to .92	.28 *(M) range:* .17 - .42	.04	.83 *(w-l) 95% CI,* .72 to .94 .83 *(w-q) 95% CI,* .71 to .96	.92 *95% CI,* .88 to .95

Note. The contingency table size represents the size of the table used to compute the reliability indexes. K represents the findings of the Kappa coefficient; (*w-q*) refers to weighted Kappa using quadratic weights, (*w-l*) refers to weighted Kappa using linear weights and kappa coefficients with *(unw)* are unweighted. 95% confidence intervals are provided for all kappa indexes. PI refers to the Prevalence Index; (*M*) represents the mean PI whenever applicable. BI refers to the Bias Index. For the PABAK coefficient, (*w-q*) and (*w-l*) refers to quadratic and linear weighted PABAK. ICC represents the intra-class correlation statistic and the 95% confidence interval.

Table 2.4: Interpretation of Kappa Based on Landis & Koch (1997)'s Scale

Kappa	Interpretation
< 0	Less than chance agreement
.01 - .20	Slight agreement
.21 - .40	Fair agreement
.41 - .60	Moderate agreement
.61 - .80	Substantial agreement
.81 - .99	Almost perfect agreement

Note. A Kappa measure of 0 equates chance while a Kappa of 1 equates a perfect agreement. Negative Kappa measures are possible and represent less agreement than chance (i.e., disagreement).

Finally, we assessed whether testing location (either in-lab or off-site) was related to knower level classification. To do so, we conducted an ordinal logistic regression ("porl" function in MASS package in R; Venables & Ripley, 2002) with knower levels as the dependent variable and location as the predictor, which revealed no significant effect of location (t = .64; p = .52).[11] We also conducted a Fisher's exact test to see if there was a difference in agreement (i.e., matches vs non-matches) between knower levels at T1 and T2 based on testing location, but this was not the case (p = .43).

Discussion

In Experiment 1, we found that the titrated Give-N task was highly reliable both when all knower levels were considered at once and when considering knower level groups (i.e., subset-knowers, non-knowers, and CP-knowers). The results using the ICC statistic corroborated these findings. However, we noted substantial variation in the concordance of individual knower levels, particularly within the group of subset-knowers, with relatively high concordance for non-

[11] This is the result obtained when knower levels at T1 are used as the dependent variable. We obtained the same outcome when using knower levelsat T2 (t = -0.06; p = .95)

knowers, 1-knowers, 2-knowers, and CP-knowers, but lower concordance for 3-, 4-, and 5-knowers.

Experiment 2: Give-a-Number Non-Titrated

In Experiment 2, we assessed the test-retest reliability of the non-titrated version of Give-N.

Method

Participants

In total, 101 English-speaking children were tested for this experiment. Twenty children were excluded because of (1) failure to complete all 3 tasks (n = 12), (2) language barrier (n = 1), (3) not being in the targeted age range (n = 5), and (4) experimenter error (n = 2), leaving a final sample of 81 children, aged 2;6 to 4;1-year-old (M = 3;4 years). Children were recruited in the same way as in Experiment 1.

Materials and procedure

The testing environments were the same as in Experiment 1, except that children were presented with a non-titrated version of Give-N, twice, separated by the Highest Count task. Because the non-titrated version of Give-N includes a fixed number of trials, each session lasted approximately 10 minutes, slightly longer than in Experiment 1.

Non-Titrated Give-a-Number Task

This task was identical to the titrated version used in Experiment 1, except for the trial structure. Children were given 15 trials including three for each of the numbers 1, 2, 3, 4, and 6. Note that since we did not ask for *five*, children could not be classified as 5-knowers in this version, unlike in the titrated task (though, in Experiment 1, only 5 children were ever classified as 5-

knowers). We created two lists of trials in a pseudorandom order. All children were presented with both lists counterbalanced in order at T1 and T2 across children. The criteria to assign knower levels were the same as those used in the titrated version, with an emphasis on the requirement for children to succeed at all numbers below N to be credited as N-knowers (i.e., children couldn't skip some numbers). Children were credited as CP-knowers if they could correctly give six on two out of three trials.

Highest Count (HC)

The task was identical to Experiment 1.

Results

Table 2.5 shows the distribution of knower levels in the first and second assessments of the task. As in Experiment 1, most children counted just above 10 in the Highest Count task (M = 13.6), and their counting skills were variable (*range* = 1 to 59; SD = 12.2). Seventy-seven children out of 81 (95%) had a highest count higher than their knower level across the two Give-N tasks.

We first calculated the reliability of the non-titrated task, including all knower levels (0 to CP) in a 6x6 contingency table.[12] All statistics are summarized in Table 2.6. We found an overall agreement of 72% and a $K_{w\text{-}linear}$ of .81 and .90 (quadratic; $Kappa_{unweighted}$ = .63; $Prevalence\ index_{(mean)}$ = .12, *range* = .03 - .28; *Bias index* = .04; $PABAK_{(weighted)}$ = .66; *ICC* = .95). Figure 2.2 illustrates the contingency table used and the knower levels at T1 and T2 as well as their effective agreement.

Next, we explored the reliability for subset-knowers, non-knowers and CP-knowers separately. All Kappas were unweighted in these analyses. For the subset-knower analysis, we created a 5x5 contingency table with knower levels 1 to 4 and a non-subset-knower category at T1 and T2. We found an effective agreement of 56% and a Kappa of .61 ($Prevalence\ index_{(mean)}$ = .16,

[12] Note that since we did not test for 5, children could not be classified as 5-knower, reducing the knower level categories from 7 (0 to CP; as in Exp 1) to 6.

range = .04 - .33; *PABAK* = .65).[13] In the non-knower analysis (2x2 contingency table), we obtained an effective agreement of 57% and a Kappa of .71 (*Prevalence index* = .86; *Bias index* = .01; *PABAK* = .93). In the CP-knower analysis, we found an effective agreement of 76% and a Kappa of .79 (*Prevalence index* = .28; *Bias index* = .02; *PABAK* = .80). So far, the results of Experiment 2 are similar to those of Experiment 1; the overall reliability of the task was high but varied across individual knower levels, especially within the group of subset-knowers.

Next, we assessed the agreement and reliability of assignment to broader knower level groups - i.e., non-knower, subset-knower, or CP-knower. Here, we found an overall agreement of 86%, and a $K_{w\text{-}linear}$ of .77 and .80 (quadratic; $Kappa_{unweighted}$ = .75; $Prevalence\ index_{(mean)}$ = .31, *range* = .20 - .46; *Bias index* = .04; *PABAK* = .80; *ICC* = .89). This suggests that children classified as subset-knowers in the first assessment were likely to remain subset-knowers in the second assessment (as were non-knowers and CP-knowers). The ICC analysis also confirmed these results.

Next, we assessed whether there was an effect of task order. As in Experiment 1, slightly more children showed a decrease in knower level from T1 to T2 (decreased n = 13; increased n = 10) but this difference was not significant (Wilcoxon rank test; W = 3330.5; p = .86). Also, whenever there was a difference of knower levels, more children had knower levels that differed by only one level (n = 17) compared to 2 (n = 4) or 3 levels (n = 2). We also investigated whether there was an effect of trial order across the two tasks (e.g., whether children provided correct responses more frequently for trials presented in the first position vs. trials presented in third position). However, there was no such effect (z = -1.11, p = .27).

[13] Note here that, since the "non-subset-knower" category can be placed arbitrarily on either side of the contingency table, the Bias Index is not valid in this analysis.

Table 2.5: Distribution of Knower Levels at the First and Second Assessments of non-titrated Give-N

Knower Levels	0K	1K	2K	3K	4K	CP
Assessment						
Time 1	6	18	10	7	12	28
Time 2	5	21	10	11	4	30

Note. In task 1, there were 6 children classified as non-knowers, 47 subset-knowers (1K to 4K) and 28 CP-knowers. In task 2, there were 5 non-knowers, 46 subset-knowers and 30 CP-knowers.

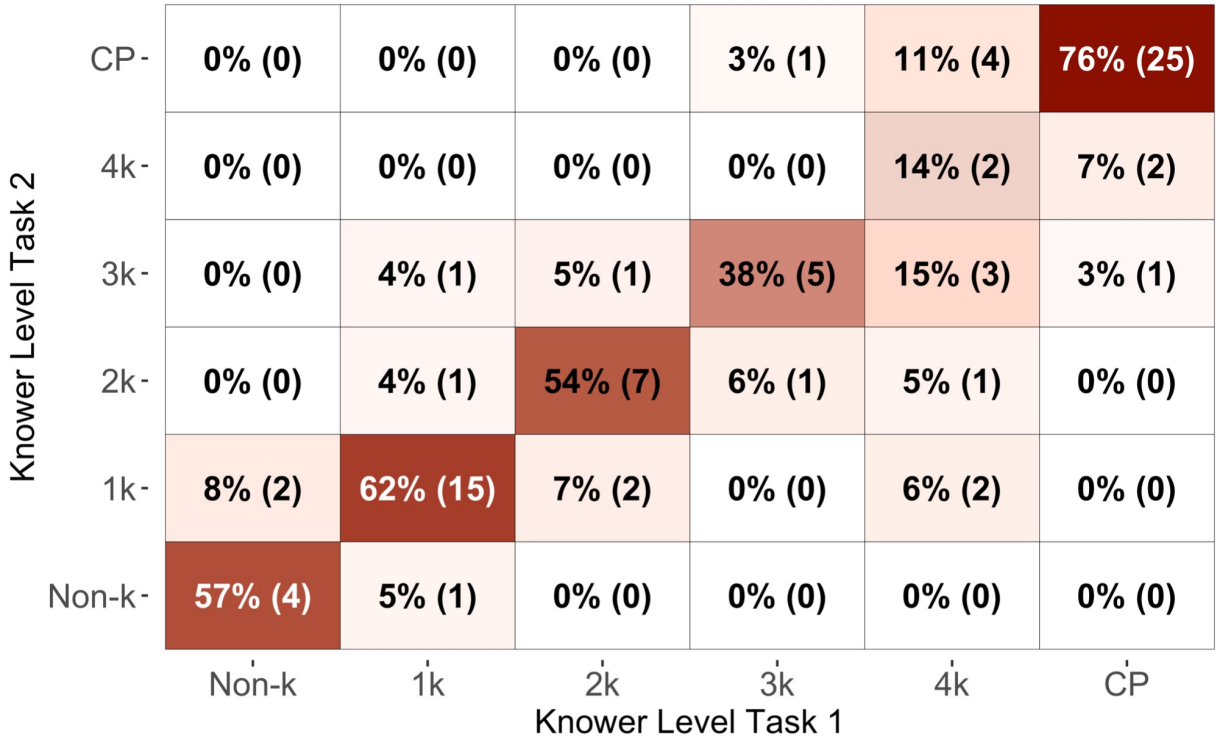

Figure 2.2: Knower level Classification in the First and Second Assessments of non-titrated Give-N

Table 2.6: Summary of Reliability measures and coefficients of the Non-Titrated Give-N at T1 and T2 across different knower levels analyses

Group	Contingency table size	Agreement	K	PI	BI	PABAK	ICC
All knower levels	6 x 6	72%	.81 *(w-l)* *95% CI,* .73 to .89 .90 *(w-q)* *95% CI,* .84 to .96 .63 *(unw)* *95% CI,* .51 to .75	.12 *(M)* *range:* .03 - .28	.04	.66 (*w-l*) *95% CI,* .52 to .80 .66 (*w-q*) *95% CI,* .48 to .84	.95 *95% CI,* .92 to .97
Subset-knower only	5 x 5	56% (effective)	.61 *(unw)* *95% CI,* .48 to .74	.16 *(M)* *range:* .04 - .33	*NA*	.65 *(unw)* *95% CI,* .52 to .77	*NA*
Non-knower vs others	2 x 2	57% (effective)	.71 *(unw)* *95% CI,* .49 to .92	.86	.01	.93 *(unw)* *95% CI,* .79 to .98	*NA*
CP-knower vs others	2 x 2	76% (effective)	.79 *(unw)* *95% CI,* .57 to 1	.28	.02	.80 *(unw)* *95% CI,* .63 to .91	*NA*
Knower-level groups	3 x 3	86%	.77 *(w-l)* *95% CI,* .64 to .90 .80 *(w-q)* *95% CI,* .69 to .91 .75 *(unw)* *95% CI,* .61 to .89	.31 *(M)* *range:* .20 - .46	.04	.80 *(w-l)* *95% CI,* .68 to .92 .80 *(w-q)* *95% CI,* .66 to .94	.89 *95% CI,* .83 to .93

Note. The contingency table size represents the size of the table used to compute the reliability indexes. K represents the findings of the Kappa coefficient; (*w-q*) refers to weighted Kappa using quadratic weights, (*w-l*) refers to weighted Kappa using linear weights and kappa coefficients with *(unw)* are unweighted. 95% confidence intervals are provided for all kappa indexes. PI refers to the Prevalence Index; (*M*) represents the mean PI whenever applicable. BI refers to the Bias Index. For the PABAK coefficient, (*w-q*) and (*w-l*) refers to quadratic and linear weighted PABAK. ICC represents the intra-class correlation statistic and the 95% confidence interval.

Finally, we found no difference in the distribution of knower levels (t = -.064; p = .95) or agreement (p = .467) depending on testing location (in-lab vs. off-site).

Discussion

Against our expectation that the non-titrated Give-N would yield more reliable outcomes, the pattern of results of Experiment 2 is similar to that found in Experiment 1. Specifically, we found that the reliability of the non-titrated Give-N task was high when considering all knower levels at once, but that there was considerable variability when looking at knower levels individually. While the reliability for non-knowers was particularly high, the concordance within the group of subset-knowers varied considerably and was higher for early subset-knowers (1- and 2-knowers) than late subset-knowers (3- and 4-knowers). To better understand how the titrated and non-titrated Give-N versions compare to each other, in Experiment 3 we asked whether they would generate the same knower level within participants. Also, we asked how performance at the two versions of Give-N compared to performance on the *What's-On-This-Card* task (Gelman, 1993; Le Corre et al., 2006).

Experiment 3: Give-a-Number Titrated, Non-Titrated and What's-On-This-Card

The results of Experiments 1 and 2 suggest that both versions of Give-N have an overall high test-retest reliability. However, this high reliability does not necessarily mean that the two versions converge on the same knower levels when tested within-subjects, since a given task can be reliable despite exhibiting bias. Given this, it is possible that one version generates higher knower levels than the other. For example, because of how knower levels are defined by Wynn's criteria, the inclusion of more trials in the non-titrated version may result in a more conservative – and therefore lower – knower level estimate. Specifically, because a child is only considered an N-knower if they give N for requests of N but not for larger numbers, the inclusion of a greater

number of trials creates greater opportunity for random error, possibly resulting in lower knower level estimates. Concretely, if a child correctly gives 3 objects when asked for *three* on both tasks, but then gives 3 objects on ¼ of remaining trials for larger numbers, this will not impact their knower level assignment when using the titrated method (which may include as few as 2 trials above their knower level). However, the knower level may be impacted when using the non-titrated method – e.g., if the child receives 3 trials for each of tested with *four, six,* and *eight,* since ¼ of 9 these trials (i.e., ~2) would constitute 50% of all trials in which 3 is given by the child, ruling out the classification of the child as a 3-knower according to the criteria described above.

While differences between the two Give-N versions can be assessed by directly comparing their outputs, another approach is to ask how each task relates to independent measures of number knowledge. Although there is no single task that tests exactly the same construct as Give-N, a closely related measure is the What's-On-This-Card task (Gelman, 1993; Le Corre et al., 2006), in which children are presented with cards depicting images of sets and are asked to report how many objects they see. In Experiment 3 we administered the What's-On-This-Card task and paired it with a within-subjects comparison of performance on the titrated and non-titrated versions of Give-N. This allowed us to test whether either version of the Give-N task was more closely related to an independent test of number word knowledge.

Method

Participants

In total, 96 English-speaking children were tested in this experiment. Twenty-one children were excluded from analysis because of (1) failure to complete all 4 tasks (n = 13), (2) not being a native speaker of English (n = 1), (3) not being in the targeted age range (n = 1), (4) experimenter error (n = 3) and (5) classifying as a 5-knower in the titrated Give-N (n = 3), leaving a final sample

of 75 children, aged 2;1 to 4;0-year-old (M = 3;2 years). Because there was no difference between testing sites in Experiments 1 and 2, all children in this experiment were recruited off-site (i.e., preschools and museums).

Materials and procedure

Each session lasted approximately 15 minutes and included (1) Give-a-Number task 1, (2) Highest Count task, (3) Give-a-Number task 2, and (4) What's-On-This-Card task. All participants were administered the tasks in this order, but the order of the titrated and non-titrated Give-N tasks was counterbalanced across children. The procedures for the titrated and non-titrated tasks were identical to what is reported in Experiments 1 and 2, as were the procedures for the Highest Count task.

What's-On-This-Card (WOC)

This task was modeled after Le Corre et al. (2006). Children were presented with 15 cards containing either 1, 2, 3, 4, or 6 items (balloon, car, dog), assessed 3 times each. Children were asked to report how many items they saw on each card in the following way: "*Now, I'm going to show you some pictures. Your job is to tell me what you see in these pictures. How many [item(s)] do you see in this picture?*". After this initial question, if children did not spontaneously count the items 2 to 6, they were prompted to do so ("*Can you count them for me?"*). Items were aligned and displayed in either one or two rows (depending on the number) and varied in color to maintain children's interest. Two lists of trials in a pseudorandom order were randomly assigned to participants. Before the 15 critical trials, children were presented with a practice trial which was intended to model the expected response and encourage children to provide a number word. We used the same criteria as in Experiments 1 and 2 to assign knower levels; namely that children needed to provide correctly N two out of three times when asked for N, and do so only for N and

numbers below. Children were credited as CP-knowers if they could correctly give six on two out of three trials.

Results

We first assessed the relatedness of the titrated and non-titrated Give-N measures. Although comparing performance on these two versions of Give-N does not strictly amount to assessing reliability - since they are not the same measure - we nevertheless used the statistical tools introduced in Experiments 1 and 2 since measures of reliability offer the best way to assess how often two tasks exhibit agreement in this context. Table 2.7 shows the distribution of knower levels in the titrated and non-titrated Give-N tasks and for the WOC. On average participants could count to around 9 in the Highest Count task ($M = 9.1$) and their counting skills were variable (*range* = 0 to 30; $SD = 6.0$). Seventy children (93%) reached a knower level higher than their highest count across all 3 tasks.

Table 2.7: Distribution of Knower Levels for titrated Give-N, non-titrated Give-N and What's-On-This-Card

Knower Levels	0K	1K	2K	3K	4K	CP
Assessment			Number of Participants			
Titrated Give-N	7	20	17	10	5	16
Non-Titrated Give-N	10	25	13	8	6	13
What's-On-This-Card	5	23	12	5	11	19

Note. In titrated Give-N, there were 7 children classified as non-knowers, 52 subset-knowers and 16 CP-knowers. In non-titrated Give-N, there were 10 non-knowers, 52 subset-knowers and 13 CP-knowers. In What's-On-This-Card, there were 5 non-knowers, 51 subset-knowers and 19 CP-knowers.

Comparing titrated versus non-titrated Give-N

Table 2.8 provides a summary of the various coefficients assessing the relatedness of the titrated and non-titrated Give-N tasks. The analysis including all knower levels (0 to CP; Figure

2.3) found that the relatedness of the two Give-N versions is high but lower than the relatedness of each respective version of the task to itself (in Experiments 1 and 2), suggesting that there are some real differences between the two versions above, beyond noise associated with test-retest reliability. As shown in Figure 2.3, the degree of concordance between the tasks varies substantially across the individual knower levels. In particular, the concordance of 1-knowers and CP-knowers is especially high relative to other knower levels, resembling the reliability findings for Experiments 1 and 2.

Table 2.8: Reliability measures and coefficients between the Titrated and Non-Titrated Give-N across different knower levels analyses

Group	Contingency table size	Agreement	K	PI	BI	PABAK	ICC
All knower levels	6 x 6	59%	.70 *(w-l)* *95% CI,* .60 to .80 .84 *(w-q)* *95% CI,* .76 to .92 .49 *(unw)* *95% CI,* .35 to .62	.08 *(M)* *range:* 0 - .16	.17	.50 (*w-l*) *95% CI,* .34 to .67 .50 (*w-q*) *95% CI,* .29 to .72	.91 *95% CI,* .86 to .95
Subset-knower only	5 x 5	47% (effective)	.46 *(unw)* *95% CI,* .31 to .60	.10 *(M)* *range:* .03 - .19	*NA*	.48 *(unw)* *95% CI,* .34 to .62	*NA*
Non-knower vs others	2 x 2	31% (effective)	.41 *(unw)* *95% CI,* .18 to 63	.77	.04	.76 *(unw)* *95% CI,* .57 to .89	*NA*
CP-knower vs others	2 x 2	71% (effective)	.79 *(unw)* *95% CI,* .56 to 1	.61	.04	.87 *(unw)* *95% CI,* .70 to .96	*NA*
Knower-level groups	3 x 3	81%	.64 *(w-l)* *95% CI,* .47 to .81 .69 *(w-q)* *95% CI,* .54 to .84 .60 *(unw)* *95% CI,* .42 to .79	.36 *(M)* *range:* .11 - .55	.08	.72 *(w-l)* *95% CI,* .58 to .86 .72 *(w-q)* *95% CI,* .55 to .89	.82 *95% CI,* .71 to .89

Note. The contingency table size represents the size of the table used to compute the reliability indexes. K represents the findings of the Kappa coefficient; (*w-q*) refers to weighted Kappa using quadratic weights, (*w-l*) refers to weighted Kappa using linear weights and kappa coefficients with *(unw)* are unweighted. 95% confidence intervals are provided for all kappa indexes. PI refers to the Prevalence Index; (*M*) represents the mean PI whenever applicable. BI refers to the Bias Index. For the PABAK coefficient, (*w-q*) and (*w-l*) refers to quadratic and linear weighted PABAK. ICC represents the intra-class correlation statistic and the 95% confidence interval.

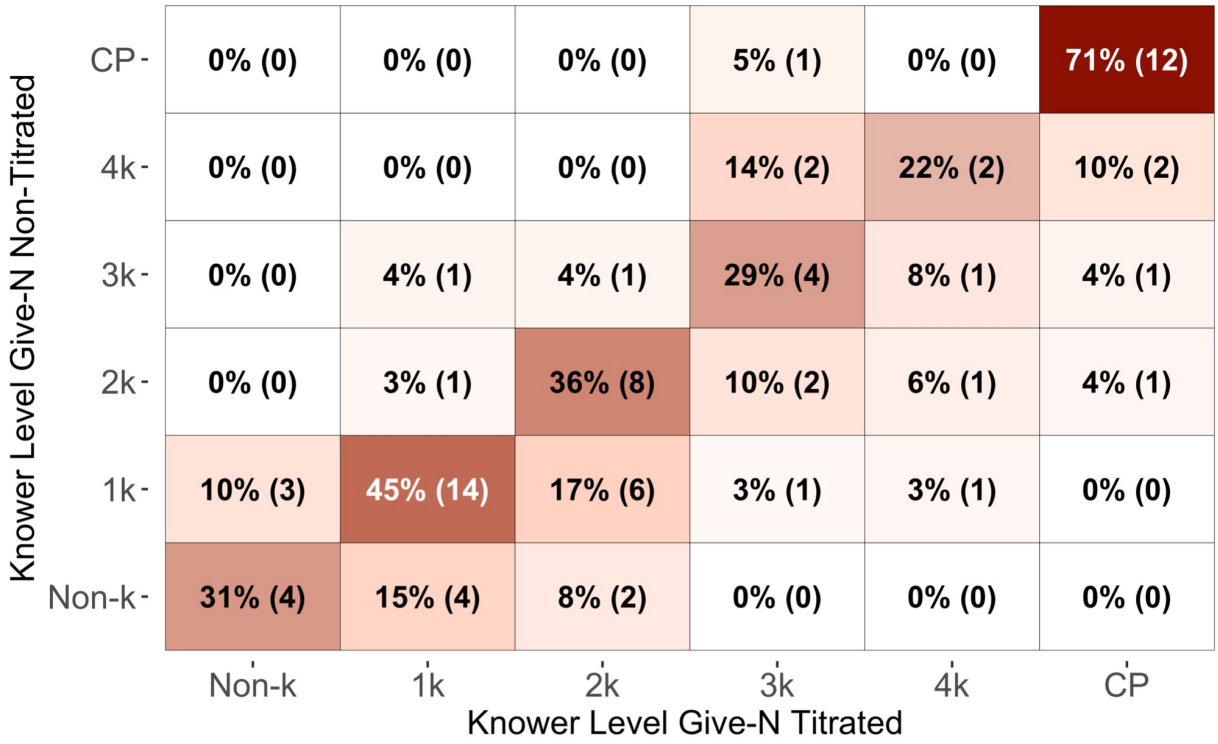

Figure 2.3: Knower level Classification for Titrated and Non-Titrated Give-N

Note. Titrated Give-N appears on the x axis, and non-titrated Give-N appears on the y axis. The percentages represent the percent effective agreement of both knower level assignments. Numbers in parentheses represent the frequency of the paired knower level. The color scale is based on the proportion of effective agreement.

To better understand how the two Give-N versions compared to each other, we next examined the differences in their outcomes (see Figure 2.4). In total, there were 44 matches (59%) and 31 non-matches (41%). Overall, the non-titrated version generated significantly lower knower levels than the titrated version (Wilcoxon test rank test $V = 132.5$; $p = .02$). However, the majority of non-matches were differences of only one knower level ($n = 22$) as opposed to 2 knower levels ($n = 7$) or 3 knower levels ($n = 2$). Furthermore, when taking into account the Kappa statistics presented earlier, it appears that most children with non-matches were subset-knowers in both Give-N versions.

Next, we investigated whether there was an order effect. To do this, we performed a Wilcoxon rank test and found no significant effect of task order ($W = 2893$, $p = .76$). Regardless of Give-N type, from T1 to T2, a similar number of children increased their knower levels (n = 17) as decreased (n = 14).

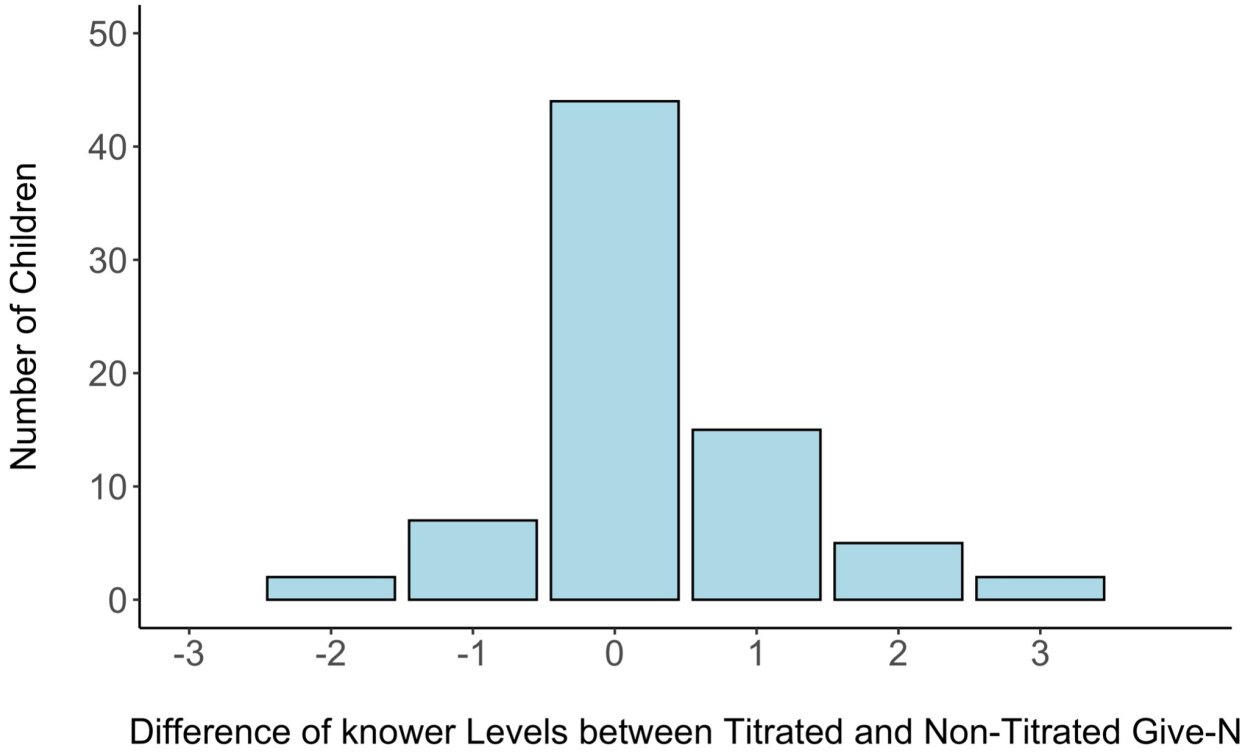

Figure 2.4: Differences in Participant Knower Level Across Give-N Versions

Note. The x-axis refers to the difference in participant knower level assignment between the titrated and non-titrated versions of Give-N. The 0 indicates no change in knower level assignment across the versions, while a positive number indicates a higher knower level assignment for titrated Give-N and a negative number indicates a higher knower level assignment for non-titrated Give-N. Amongst the 31 children with no matches between knower level assignments, 22 children had a higher knower level in the titrated version while 9 children had a higher knower level in the non-titrated version.

Comparing the Knower Levels of Give-N Titrated and WOC

Table 2.9 provides a summary of the various coefficients assessing the relatedness between the titrated Give-N and WOC. The analysis including all knower levels (0 to CP; Figure 2.5) found that the relatedness of the two tasks is acceptable but much lower than the relatedness between the

two Give-N versions reported above. The ICC statistics corroborate those results. Figure 2.5 also shows that concordance is high for 1-knowers and CP-knowers but very weak for 3- and 4-knowers. Figure 2.6 shows the distribution of differences in knower levels between WOC and titrated Give-N. Overall, slightly more children were credited with a higher knower level in WOC (n = 25) compared to titrated Give-N (n = 17), though this difference was not significant (Wilcoxon test rank test $V = 325.5$; $p = .10$). Also, whenever there was a difference, most children presented a difference of 1 knower level (n = 27), as opposed to a difference of 2 (n = 8) or more levels (3 or 4 levels; n = 7).

Table 2.9: Reliability measures and coefficients between the Titrated Give-N and WOC across different knower levels analyses

Group	Contingency table size	Agreement	K	PI	BI	PABAK	ICC
All knower levels	6 x 6	44%	.55 *(w-l)* *95% CI,* .43 to .67 .71 *(w-q)* *95% CI,* .58 to .83 .30 *(unw)* *95% CI,* .18 to .43	.09 *(M)* *range:* .01 - .19	.11	.33 (*w-l*) *95% CI,* .16 to .50 .33 (*w-q*) *95% CI,* .10 to .56	.83 *95% CI,* .73 to .89
Subset-knower only	5 x 5	32% (effective)	.27 *(unw)* *95% CI,* .14 to .40	.11 *(M)* *range:* .01 - .19	*NA*	.30 *(unw)* *95% CI,* .16 to .44	*NA*
Non-knower vs others	2 x 2	20% (effective)	.28 *(unw)* *95% CI,* -.08 to .64	.84	.03	.79 *(unw)* *95% CI,* .60 to .91	*NA*
CP-knower vs others	2 x 2	46% (effective)	.52 *(unw)* *95% CI,* .29 to .74	.53	.04	.65 *(unw)* *95% CI,* .44 to .81	*NA*
Knower-level groups	3 x 3	72%	.45 *(w-l)* *95% CI,* .25 to .64 .52 *(w-q)* *95% CI,* .35 to .69 .40 *(unw)* *95% CI,* .20 to .61	.35 *(M)* *range:* .12 - .52	.07	.58 *(w-l)* *95% CI,* .42 to .74 .58 *(w-q)* *95% CI,* .38 to .78	.69 *95% CI,* .50 to .80

Note. The contingency table size represents the size of the table used to compute the reliability indexes. K represents the findings of the Kappa coefficient; (*w-q*) refers to weighted Kappa using quadratic weights, (*w-l*) refers to weighted Kappa using linear weights and kappa coefficients with *(unw)* are unweighted. 95% confidence intervals are provided for all kappa indexes. PI refers to the Prevalence Index; (*M*) represents the mean PI whenever applicable. BI refers to the Bias Index. For the PABAK coefficient, (*w-q*) and (*w-l*) refers to quadratic and linear weighted PABAK. ICC represents the intra-class correlation statistic and the 95% confidence interval.

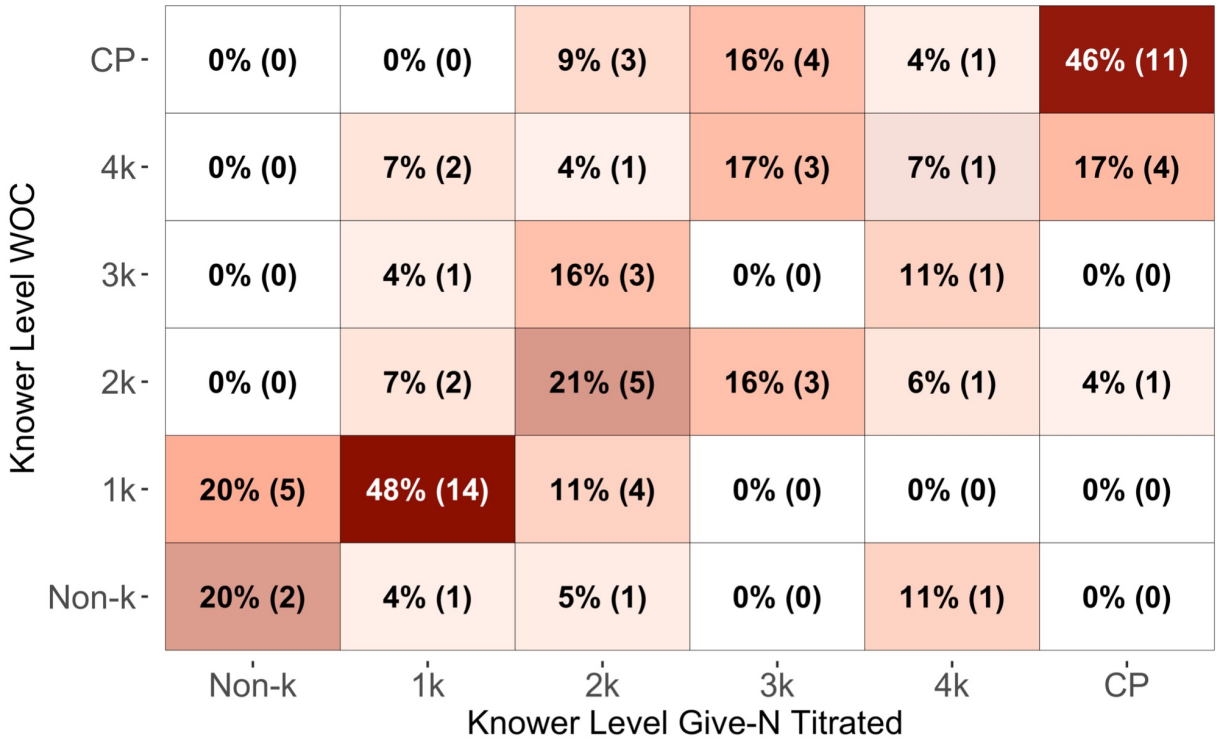

Figure 2.5: Knower level Classification for Titrated Give-N and WOC

Note. Titrated Give-N appears on the x axis, and WOC appears on the y axis. The percentages represent the percent effective agreement of both knower level assignments. Numbers in parentheses represent the frequency of the paired knower level. The color scale is based on the proportion of effective agreement.

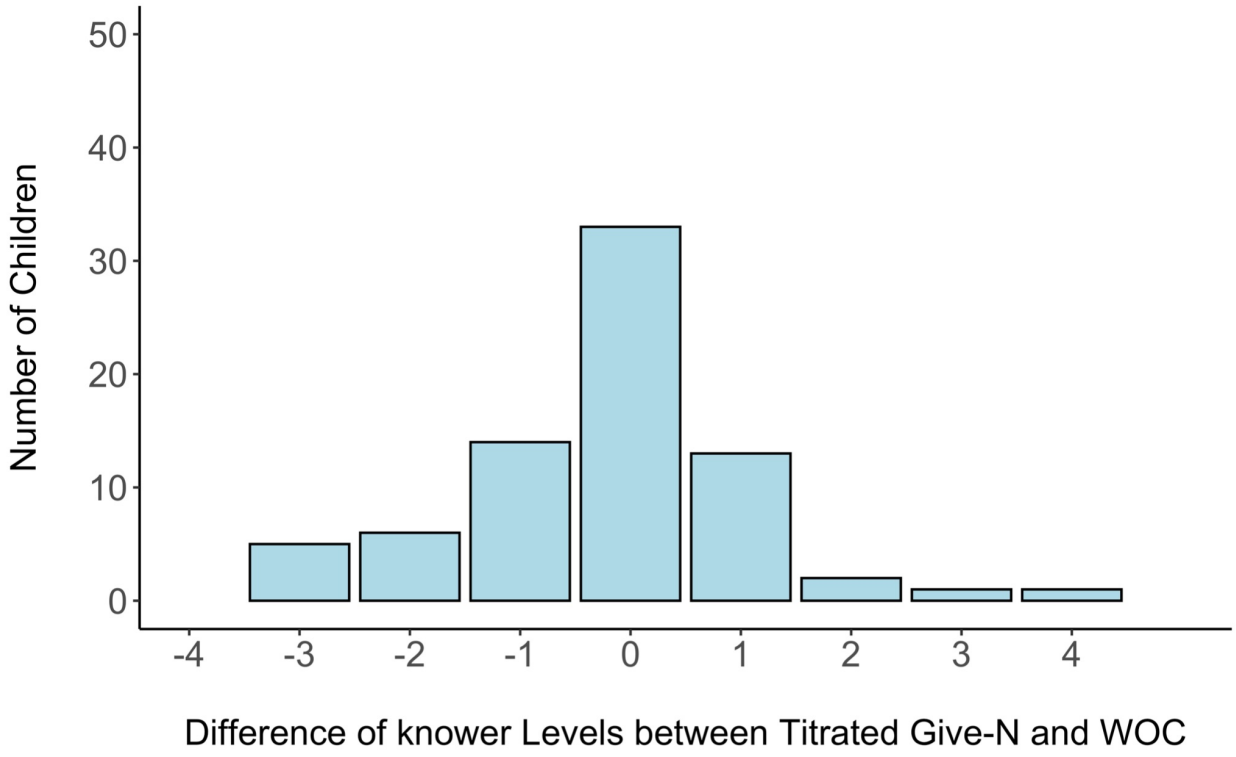

Figure 2.6: Differences in Participant Knower Level Between Titrated Give-N and What's-On-This-Card

Note. The x-axis refers to the difference in participant knower level assignment between the titrated version of Give-N and the What's-On-This-Card task such that 0 indicates no change in knower level assignment across the tasks, a positive number indicates a higher knower level assignment for titrated Give-N, and a negative number indicates a higher knower level assignment for What's-On-This-Card. There were 33 children with matching knower levels in the two tasks. Amongst the 42 children with no-matches between knower level assignments, 25 had a higher knower level in WOC and 17 in titrated Give-N.

Comparing the Knower Levels of Non-Titrated Give-N and WOC

Table 2.10 summarizes the various coefficients assessing the relatedness between the non-titrated Give-N and WOC. The analysis including all knower levels (0 to CP; Figure 2.7) found that the relatedness of the two tasks is acceptable, similar to the outcome presented above of the assessment of WOC and titrated Give-N. This was also the case for the ICC statistic. The results are also similar in that the concordance was strongest for 1-knowers and CP-knowers but weaker for other knower levels. Figure 2.8 shows that overall, more children were credited with a higher

knower level in WOC (n = 27) compared to non-titrated Give-N (n = 11) and this difference was significant (Wilcoxon test rank test $V = 172$; $p = .003$). Interestingly, unlike in the case of the titrated task, the differences in knower levels between WOC and the non-titrated task tended to be larger; 21 children had knower levels that differed by one level while 17 participants had knower levels that differed by 2 levels (n = 8) or more (3 levels n = 7; 4 or 5 levels n = 2).

Table 2.10: Reliability measures and coefficients between the Non-Titrated Give-N and WOC across different knower levels analyses

Group	Contingency table size	Agreement	K	PI	BI	PABAK	ICC
All knower levels	6 x 6	49%	.54 *(w-l)* *95% CI,* .41 to .67 .66 *(w-q)* *95% CI,* .51 to .81 .37 *(unw)* *95% CI,* .23 to .50	.08 *(M)* *range:* 0 - .19	.21	.39 (*w-l*) *95% CI,* .22 to .56 .39 (*w-q*) *95% CI,* .17 to .62	.80 *95% CI,* .65 to .88
Subset-knower only	5 x 5	40% (effective)	.34 *(unw)* *95% CI,* .20 to .49	.10 *(M)* *range:* .03 - .19	*NA*	.38 *(unw)* *95% CI,* .24 to .52	*NA*
Non-knower vs others	2 x 2	25% (effective)	.34 *(unw)* *95% CI,* .13 to .55	.80	.07	.76 *(unw)* *95% CI,* .57 to .89	*NA*
CP-knower vs others	2 x 2	39% (effective)	.45 *(unw)* *95% CI,* .23 to .67	.57	.08	.63 *(unw)* *95% CI,* .41 to .79	*NA*
Knower-level groups	3 x 3	71%	.41 *(w-l)* *95% CI,* .22 to .61 .46 *(w-q)* *95% CI,* .25 to .66 .38 *(unw)* *95% CI,* .18 to .59	.34 *(M)* *range:* .08 - .51	.13	.56 *(w-l)* *95% CI,* .39 to .73 .56 *(w-q)* *95% CI,* .36 to .76	.63 *95% CI,* .41 to .77

Note. The contingency table size represents the size of the table used to compute the reliability indexes. K represents the findings of the Kappa coefficient; (*w-q*) refers to weighted Kappa using quadratic weights, (*w-l*) refers to weighted Kappa using linear weights and kappa coefficients with *(unw)* are unweighted. 95% confidence intervals are provided for all kappa indexes. PI refers to the Prevalence Index; (*M*) represents the mean PI whenever applicable. BI refers to the Bias Index. For the PABAK coefficient, (*w-q*) and (*w-l*) refers to quadratic and linear weighted PABAK. ICC represents the intra-class correlation statistic and the 95% confidence interval.

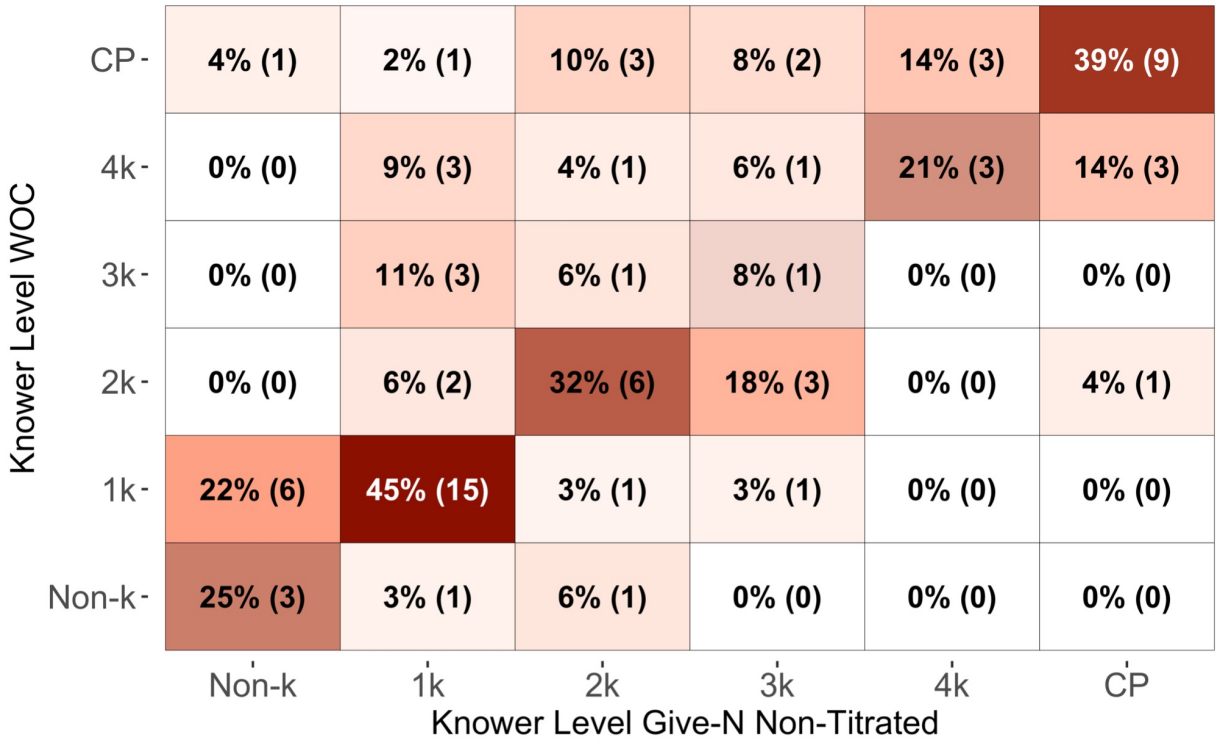

Figure 2.7: Knower level Classification for Non-Titrated Give-N and WOC

Note. Non-titrated Give-N appears on the x axis, and WOC appears on the y axis. The percentages represent the percent effective agreement of both knower level assignments. Numbers in parentheses represent the frequency of the paired knower level. The color scale is based on the proportion of effective agreement.

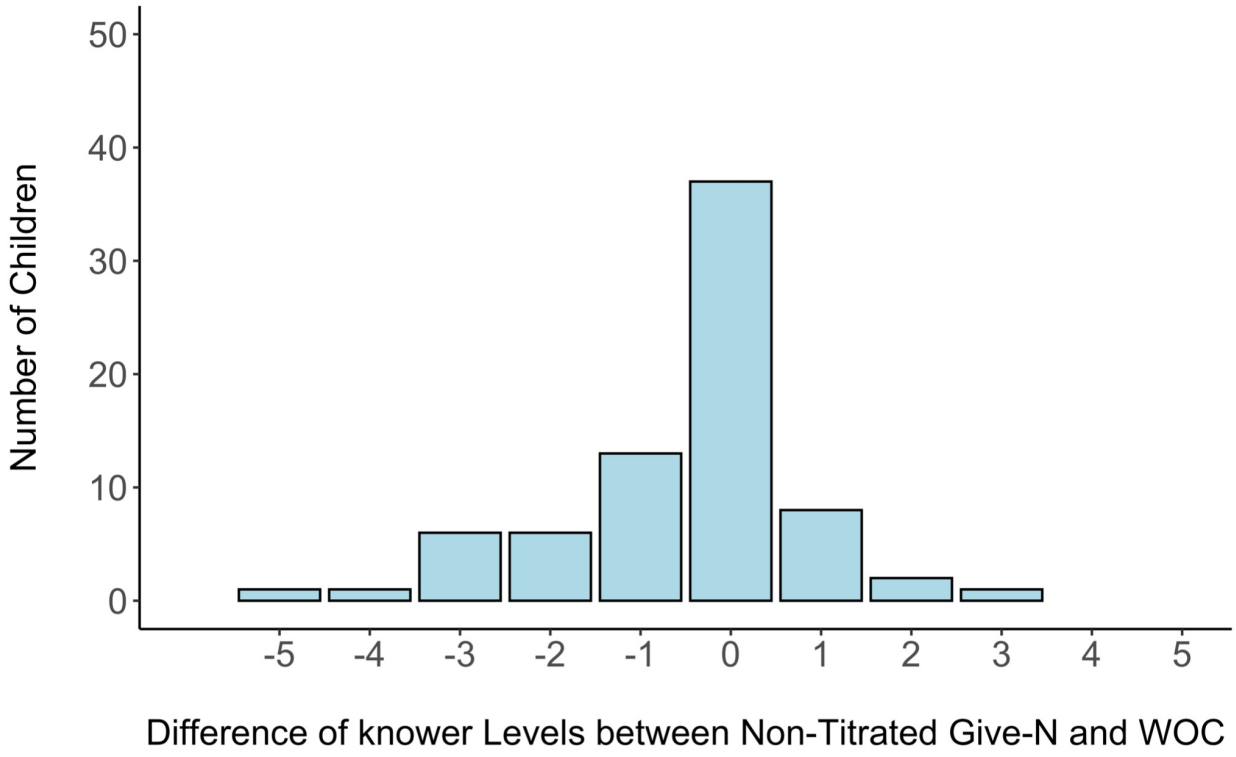

Figure 2.8: Differences in Participant Knower Level Between Non-titrated Give-N and What's-On-This-Card

Note. The x-axis refers to the difference in participant knower level assignment between the non-titrated version of Give-N and the What's-On-This-Card task such that 0 indicates no change in knower level assignment across the tasks, a positive number indicates a higher knower level assignment for non-titrated Give-N, and a negative number indicates a higher knower level assignment for What's-On-This-Card. There were 37 children who had matching knower levels across the two tasks and amongst the 38 who did not match, 27 had a higher knower level at the WOC task and 11, in the non-titrated Give-N task.

Discussion

In Experiment 3, we found that the relatedness of the titrated and non-titrated Give-N tasks was substantial, but that the titrated Give-N task generated slightly higher knower levels, typically 1 level greater than that of the non-titrated task. The comparison of WOC and the two Give-N versions found that the relatedness of WOC with either Give-N was acceptable, but not as strong as the relatedness of the two Give-N versions to one another. Interestingly, the Give-N versions differed in how they related to WOC. Although the titrated Give-N task did not generate

systematically higher or lower knower levels than WOC (differences between outcomes were random), the non-titrated version produced significantly lower knower levels than WOC. This last result is consistent with previous studies in the literature suggesting that WOC attributes more knowledge of number words to children than Give-N (Baroody et al., 2017; Mou et al., 2018; O'Rear et al., 2020).

Post Hoc Analyses

As suggested by a reviewer, we conducted *post hoc* analyses on the relationship between children's counting abilities, age, and reliability. We conducted the analyses for subset-knowers and CP-knowers separately given the qualitatively higher degree of reliability among CP-knowers and non-knowers (making linear models difficult to interpret). One possibility is that highest count influences reliability (operationalized as concordance) for both groups, if it reflects children's executive functioning, attention, or ability to learn robust representations. An alternative possibility is that highest count may predict concordance only for CP-knowers, since only they can accurately count. With respect to age, predictions are more complicated, since among subset-knowers, lower knower levels exhibit higher reliability and children with lower knower levels tend to be younger. At the same time, older children should be less variable in how they respond relative to younger children. For CP-knowers, we expected that age might be positively related to concordance (because older children are better able to regulate their behaviors), or that it might be unrelated, given that CP-knowers have uniformly high levels of reliability.

To test these possibilities, we conducted two logistic regressions predicting Concordance in knower levels (yes/no) from Age and Highest Count. To maximize statistical power, we combined data from Experiments 1, 2, and 3 (only Give-N tasks). In our first model targeting only CP-knowers, Age (β = -.21, *SD* = .10, z = -2.09, p = .04) and Highest Count (β= .07, *SD* = .03, z =

2.16, p = .03) were significant. Somewhat surprisingly, this effect of age was negative, suggesting that younger CP-knowers were slightly more likely to exhibit concordance than older CP-knowers. However, this effect was relatively small, and likely would not be found if substantially older CP-knowers were also tested. In our second model with subset-knowers (0-, 1-, 2-, 3- and 4-knowers), in which we added knower levels as a covariate, neither Age (p= .98) nor Highest Count (p = .15) were significant. Although the role of Age was not straightforward in this study, we believe that its role in influencing reliability should not be overlooked in developmental studies interested in the reliability of different tasks, as we discussed in our General Discussion. The results for Highest Count are consistent with our second prediction that counting abilities have an influence on concordance but only when children understand the purpose of counting and can use their counting skills in a task.

General Discussion

In three studies we tested the reliability of the Give-a-Number task, while comparing two commonly used versions, the titrated and non-titrated versions. Overall, we found that the Give-N task is highly reliable, regardless of which version is used, though notable differences were found both between the tasks and across individual knower levels. First, in Experiment 1 we found that the titrated version of Give-N was very reliable overall, though the concordance of individual knower levels varied considerably, such that non-knowers, 1-knowers, 2-knowers, and CP-knowers exhibited fairly high concordance, while 3-, 4-, and 5-knowers did not.[14] We also found that the task could be reliably used to assign children to a less fine-grained tripartite classification of non-knower vs. subset-knower (1- through 4-knower) vs. CP-knower. Experiment 2 found almost identical results for the non-titrated Give-N task. In both experiments, testing location

[14] Although their concordance was indeed low, 5-knowers were too infrequent (only 5 children ever obtain this classification) to draw firm conclusions from.

(either in-lab or off-site) didn't impact reliability. Finally, in Experiment 3 we tested the titrated and non-titrated versions within-subjects, and found that they exhibited a high degree of concordance, overall, although concordance was lowest for 3- and 4-knowers, similar to what was found when investigating test-retest reliability in Experiments 1 and 2. We also found that while the overall distribution of knower levels was similar across versions, the titrated version produced significantly higher knower levels than the non-titrated task, though typically by just one knower level. Finally, although both tasks revealed differences from the What's-on-this-Card task, only the non-titrated task produced differences that were systematic (i.e., non-random) in nature. Just as it produced lower outcomes relative to the titrated task, the non-titrated task also produced lower outcomes than What's-on-this-Card.

Overall, these results support the continued use of Give-a-Number as a framework for classifying children, organizing findings, and predicting outcomes on other developmental measures. Also, our findings have both practical and theoretical implications regarding the use and interpretation of Give-a-Number in future studies. These implications relate to (1) the choice of task version in different research contexts, (2) how to use number knower levels to predict other outcomes in correlational designs, and (3) the validity of the knower level framework, as it relates to both the specific knowledge that is ascribed to children at particular levels by the different versions of the task, and also the status of higher, less reliable knower levels.

First, given our finding that both versions of the Give-N task generate relatively high degrees of reliability, the choice of which version to use should not hinge on reliability, but instead on secondary concerns of experimental design. On one hand, the titrated Give-N task features fewer trials, requiring less time, and does not require children who have relatively low knower levels to needlessly complete trials for large and unfamiliar numbers. For these reasons, it may be

favored when Give-N is one of many tasks being administered, and when children are relatively young and unlikely to generate useful data for larger numbers. On the other hand, the titrated version of the task is considerably harder for inexperienced experimenters to learn and administer, since it requires adaptively changing the trial structure depending on children's individual behaviors, potentially increasing the likelihood of experimenter error. Also, because the titrated version does not systematically generate data for large numbers, it is not well suited to studies that seek to investigate how children respond to less familiar numbers (e.g., to test for knowledge beyond the child's knower level; Gunderson et al., 2015; O'Rear et al., 2020; Wagner & Johnson, 2011; Wagner et al., 2019), or that seek to conduct individual differences analyses, which generally assume that all participants have received the same measures (Geary, 2018; Geary et al., 2018, 2019; Shusterman et al. 2016, 2017).

A second implication of this study concerns the use of Give-N to predict other developmental outcomes. Given the relatively high reliability of Give-N, our results suggest that it can be used in several different ways to fruitfully predict outcomes of other experimental measures, such as later mathematics achievement.[15] As noted in the Introduction, the use of a measure like Give-N to meaningfully predict other variables depends upon a relatively high test-retest reliability, since the strength of a correlation between any two variables is limited by the size of the correlation between the true value of the variables being measured, and the test-retest reliability of these measures taken individually. Therefore, in a study that attempts to correlate number knower level with another measure – e.g., a child's accuracy when making numerical estimates of dot arrays – the largest reliable correlation we might find between these measures

[15] Note that in our study we have no evidence that the titrated or non-titrated version of Give-N would be more related to later learning, though other studies suggest that the non-titrated version may be more sensitive to small differences between children (e.g., O'Rear et al., 2020).

would be limited by the reliability of Give-N (around .7) and the reliability of estimation accuracy (which is somewhat lower, around .57; Inglis & Gilmore, 2014). In this example, if the true correlation between these outcomes were 100%, then the highest detectable correlation would be .63 (i.e., 1 x √(.7 x .57)). This, in turn, has implications for the power required to detect reliable correlations, and thus for the size of the sample required for the study.

The third main implication of this study relates to the validity of the knower level system, and how individual knower level assignments should be interpreted. Across different studies using the Give-N task, researchers have often assumed, following Wynn (1992), that there are roughly five categories into which children might be classified - i.e., non-knowers, 1-knowers, 2-knowers, 3-knowers, and CP-knowers. However, some have allowed for the identification of higher levels, including 4-knowers and 5-knowers, and in some cases even higher. This approach is understandable, since it is possible that by restricting the possible subset categories to just three levels, researchers may underestimate the associative meanings that children acquire before they learn to accurately count and give large sets (and become CP-knowers). Our study, however, draws into question the interpretation of these higher knower levels. As we showed across three studies, whereas the non-knower, CP-knower, 1-knower, and 2-knower stages each individually exhibit high test-retest concordance, the 3-, 4-, and 5-knower stages are substantially less stable across sessions.

This apparent instability of higher knower levels is compatible with several interpretations. One possibility is that children at these higher knower levels are not actually 4- or 5-knowers, but instead are CP-knowers who attempt to count and make errors. Compatible with this, when we look at the three children from Experiment 3 who were classified as 5-knowers in the titrated Give-N, two of them were classified as CP-knowers in the non-titrated Give-N and one of them became

a 4-knower. Similarly, one recent study by Krajcsi (2021) found that when children were prompted to fix Give-N errors by counting, this resulted in significantly more CP-knowers than when children were not prompted, or simply asked, "Is that N?" (see Le Corre et al., 2006, for related evidence). A second possibility is that children at higher subset levels aren't misclassified CP-knowers, but instead have noisy associative mappings between number words and approximate magnitudes. While some studies have argued for such a possibility (e.g., Wagner & Johnson, 2011), others have pointed out that such evidence is not robust once knower levels are assigned in keeping with Wynn's criteria, and when only those numbers clearly beyond the child's knower level are considered (e.g., Knower Level +1; see Barner & Bachrach, 2010; Gunderson et al., 2015; Wagner et al., 2019; O'Rear et al., 2020). For these reasons, it is important in future work to not only assess whether children respond correctly on initial Give-N trials, but also (1) whether their initial response was the result of "grabbing" sets or an erroneous count (see Wynn, 1992), (2) whether they are able to fix their responses via counting when prompted, and (3) whether their overall pattern of responses for larger numbers is compatible with approximation, counting, or randomly guessing. Meanwhile, however, there is strong evidence that many children with higher knower levels are simply rare misclassifications of CP-knowers, and that when children are guessing noisily, this is restricted to the small number range (i.e., sets of 3-4 or less). Our work suggests that if children can be classified into higher subset-knower levels, these classifications are not reliable and should therefore be interpreted with caution. Future research should further explore this issue, and how the use of Give-N to identify higher subset stages might be validated.

The current study has several limitations that might be addressed in future studies. First, as in many studies of the Give-N task, the inferences permitted by our study is limited by sample size, which can impact estimates of reliability (Sim & Wright, 2005; Shoukri et al., 2004). Ideally,

in order to perform fine grained analysis of different subset-knower levels, one would want large numbers of participants categorized in each knower-level. However, because of their relatively low frequency, late subset-knowers can be particularly difficult to identify. For example, to obtain just 50 3-knowers, at a liberal rate of 8 per 100 children (based on our sample from Experiment 1), at least 500 children would need to be tested. Future studies might address this problem by combining the data collection efforts of multiple labs. A second potential limitation of this study is that sample characteristics (age ranges, cultural groups, socioeconomic groups, etc.) may impact reliability, leaving open the possibility that reliability may differ in different groups. For example, targeting children who progress through the knower level stages at a later age might result in higher reliability, if older children exhibit fewer random errors in performance. Similarly, the reliability of the CP- stage may be lower in cultures where children receive less training on counting routines than in the US (e.g., see Almoammer et al., 2013). Future studies should not assume that reliability will be identical across samples with different characteristics. A third limitation of our study is that we did not manipulate the time interval between the two Give-N tasks. For practical reasons, the Give-N tasks were administered in the same testing session, with only a brief counting task between administrations. Although we didn't find evidence for significant order effects, it is possible that reliability would be even greater with longer delays, given that the evidence for fatigue effects was slightly greater than evidence for improvement over the two sessions (in Experiment 1 and 2).

In summary, we found evidence that the Give-N task provides a useful framework for classifying children's number knowledge, and that it can be fruitfully used to explore correlations with other robust developmental phenomena. It will be important for future research to explore the impact of these findings on previously published work and to systematically examine the reliability

of other tasks (e.g., WOC and Highest Count) frequently used in the literature. Given the widespread use of Give-N in the literature, future studies should also investigate the status of less reliable knower level stages, and their significance to theories of number word learning.

Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in Marchand, E., Lovelett, J. T., Kendro, K., and Barner, D. (2022). Assessing the knower-level framework: How reliable is the Give-a-Number task?. *Cognition*, 222, 104998. The dissertation author was the primary investigator and author of this paper.

This work received support from the Social Sciences and Humanities Research Council of Canada via a fellowship to E.M., a James S. McDonnell Foundation award to D.B., and a National Science Foundation award (ID: 2000827) to D.B. We would like to thank the participating children and families from Adventure Days Preschool, Seaside Preschool, Bright Beginnings Preschool, Chase Ranch Preschool, Gillispie School, Ridge City Preschool, Shelly Izzo's Daycare, St. Michael's Preschool, Northminster Preschool, Birch Aquarium, and Fleet Science Center. Special thanks as well to Samuel Beech, Anna Duran, Chris Fernandez, Hortesia Flores, Sonora Grimsted, Sara Lee, Emily Liu, Samuel Lucero, Ashlie Pankonin, and Kaithlyn Seifert. We also thank the members of the Language and Development Lab at UCSD, Attila Krajcsi and anonymous reviewers for their helpful feedback on this work.

References

Almoammer, A., Sullivan, J., Donlan, C., Marušič, F., O'Donnell, T., & Barner, D. (2013). Grammatical morphology as a source of early number word meanings. *Proceedings of the National Academy of Sciences*, *110*(46), 18448-18453.

Abreu-Mendoza, R. A., Soto-Alba, E. E., & Arias-Trejo, N. (2013). Area vs. density: influence of visual variables and cardinality knowledge in early number comparison. *Frontiers in Psychology*, *4*, 805.

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology, 60*(1), 40-62.

Barner, D., Chow, K., & Yang, S. J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology*, *58*(2), 195-219.

Barner, D., Libenson, A., Cheung, P., & Takasaki, M. (2009). Cross-linguistic relations between quantifiers and numerals in language acquisition: Evidence from Japanese. *Journal of Experimental Child Psychology, 103*(4), 421-440.

Baroody, A. J., Lai, M. L., & Mix, K. S. (2017). Assessing early cardinal-number concepts. In *Proceedings for the Thirty-ninth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (p. 324).

Buelow, M. (2020). *Risky decision making in psychological disorders*. Academic Press.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423-429.

Carey, S., & Barner, D. (2019). Ontogenetic origins of human integer representations. *Trends in Cognitive Sciences*, *23*(10), 823-835.

Ceylan, M., & Aslan, D. (2018). Cardinal Number Acquisition of Turkish Children. *Journal of Education and e-Learning Research*, *5*(4), 217-224.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of cCinical Epidemiology*, *43*(6), 551-558.

Chu, F. W., vanMarle, K., & Geary, D. C. (2016). Predicting children's reading and mathematics achievement from early quantitative knowledge and domain-general cognitive abilities. *Frontiers in Psychology*, *7*, 775.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

Condry, K. F., & Spelke, E. S. (2008). The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology: General*, *137*(1), 22.

Davidson, K., Eng, K., & Barner, D. (2012). Does learning to count involve a semantic induction?. *Cognition*, *123*(1), 162-173.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543-549.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions, Wiley-Interscience. Hoboken, NJ.

Geary, D. C. (2018). Growth of symbolic number knowledge accelerates after children understand cardinality. *Cognition*, *177*, 69-78.

Geary, D. C., & Vanmarle, K. (2016). Young children's core symbolic and nonsymbolic quantitative knowledge in the prediction of later mathematics achievement. *Developmental Psychology, 52*(12), 2130.

Geary, D. C., Vanmarle, K., Chu, F. W., Hoard, M. K., & Nugent, L. (2019). Predicting age of becoming a cardinal principle knower. *Journal of Educational Psychology*, *111*(2), 256.

Geary, D. C., vanMarle, K., Chu, F. W., Rouder, J., Hoard, M. K., & Nugent, L. (2018). Early conceptual understanding of cardinality predicts superior school-entry number-system knowledge. *Psychological Science, 29*(2), 191-205.

Gelman, R. (1993). A rational-constructivist account of early learning about numbers and objects. *Learning and Motivation*, *30*, 61-96.

Gunderson, E. A., Spaepen, E., & Levine, S. C. (2015). Approximate number word knowledge before the cardinal principle. *Journal of Experimental Child Psychology, 130,* 35-55.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23.

Inglis, M., & Gilmore, C. (2014). Indexing the approximate number system. *Acta psychologica*, *145*, 147-155.

Jara-Ettinger, J., Piantadosi, S., Spelke, E. S., Levy, R., & Gibson, E. (2017). Mastery of the logic of natural numbers is not the result of mastery of counting: Evidence from late counters. *Developmental Science, 20*(6), e12459.

Krajcsi, A. (2021). Follow-up questions influence the measured number knowledge in the Give-a-number task. *Cognitive Development*, *57*, 100968.

Krajcsi, A., Fintor, E., & Hodossy, L. (2018). A refined description of preschoolers' initial symbolic number learning. https://doi.org/10.31219/osf.io/2kh9s

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Le Corre, M. (2014). Children acquire the later-greater principle after the cardinal principle. *British Journal of Developmental Psychology*, *32*(2), 163-177.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395-438.

Le Corre, M., Li, P., Huang, B. H., Jia, G., & Carey, S. (2016). Numerical morphology supports early number word learning: Evidence from a comparison of young Mandarin and English learners. *Cognitive Psychology*, *88*, 162-186.

Le Corre, M., Van de Walle, G., Brannon, E. M., & Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cognitive Psychology, 52*(2), 130-169.

Li, P., Le Corre, M., Shui, R., Jia, G., & Carey, S. (2003). Effects of plural syntax on number word learning: A cross-linguistic study. In *28th Boston University Conference on Language Development*, Boston, MA.

Luck, S. J., Cooper, H., Camic, P., Long, D., Panter, A., Rindskopf, D., & Sher, K. (2012). *APA Handbook of Research Methods in Psychology: Volume 1, Foundations, Planning, Measures, and Psychometrics*.

Marchand & Barner. (2019). The Acquisition of French Un. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Marušič, F., Žaucer, R., Plesničar, V., Razboršek, T., Sullivan, J., & Barner, D. (2016). Does grammatical structure accelerate number word learning? Evidence from learners of dual and non-dual dialects of Slovenian. *PloS One*, *11*(8), e0159208.

Meyer, C., Barbiers, S., & Weerman, F. (2020). Many systems, one strategy: Acquiring ordinals in Dutch and English. *Glossa: A Journal of General Linguistics*, *5*(1).

Meyer, D., Zeileis, A., & Hornik, K. (2021). *vcd: Visualizing Categorical Data*. R package version 1.4-9.

Moore, A. M., VanMarle, K., & Geary, D. C. (2016). Kindergartners' fluent processing of symbolic numerical magnitude is predicted by their cardinal knowledge and implicit understanding of arithmetic 2 years earlier. *Journal of Experimental Child Psychology*, *150*, 31-47.

Mou, Y., Berteletti, I., & Hyde, D. C. (2018). What counts in preschool number knowledge? A Bayes factor analytic approach toward theoretical model development. *Journal of Experimental Child Psychology*, *166*, 116-133.

Mussolin, C., Nys, J., Content, A., & Leybaert, J. (2014). Symbolic number abilities predict later approximate number system acuity in preschool children. *PLoS One*, *9*(3), e91839.

Negen, J., & Sarnecka, B. W. (2012). Number-concept acquisition and general vocabulary development. *Child Development*, *83*(6), 2019-2027.

Newman, A., Dickstein, R., & Gargan, M. (1978). Developmental effects in social facilitation and in being a model. *The Journal of Psychology*, *99*(2), 143-150.

Nikoloska, A. (2009). Development of the cardinality principle in Macedonian preschool children. *Psihologija*, 42(4), 459-475.

Nunnally, J.C. (1970). Introduction to psychological measurement. New York: McGraw-Hill.

O'Rear, C. D., McNeil, N. M., & Kirkland, P. K. (2020). Partial knowledge in the development of number word understanding. *Developmental science*, *23*(5), e12944.

Pfefferle, J. C., Machen, J. B., Fields, H. W., & Posnick, W. R. (1982). Child behavior in the dental setting relative to parental presence. *Pediatric Dentistry*, *4*(4), 311-6.

Piantadosi, S. T., Jara-Ettinger, J., & Gibson, E. (2014). Children's learning of number words in an indigenous farming-foraging group. *Developmental Science*, *17*(4), 553-563.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199-217.

Purpura, D. J., & Simms, V. (2018). Approximate number system development in preschool: What factors predict change?. *Cognitive Development*, *45*, 31-39.

R Core Team (2018). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http s. *www. R-proje ct. org*.

Rasmussen, E. E., Keene, J. R., Berke, C. K., Densley, R. L., & Loof, T. (2017). Explaining parental coviewing: The role of social facilitation and arousal. *Communication Monographs*, *84*(3), 365-384.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition, 108*(3), 662-674.

Sarnecka, B. W., & Lee, M. D. (2009). Levels of number knowledge in early childhood. *Journal of Experimental Child Psychology, 103*(3), 325–337.

Sarnecka, B. W., Kamenskaya, V. G., Yamana, Y., Ogura, T., & Yudovina, Y. B. (2007). From grammatical number to exact numbers: Early meanings of ‘one’,‘two’, and ‘three’ in English, Russian, and Japanese. *Cognitive Psychology*, *55*(2), 136-168.

Sarnecka, B. W., Negen, J., & Goldman, M. C. (2018). Early number knowledge in dual-language learners from low-SES households. In *Language and culture in mathematical cognition* (pp. 197-228). Academic Press.

Sarnecka, B. W., & Wright, C. E. (2013). The idea of an exact number: Children's understanding of cardinality and equinumerosity. *Cognitive science*, *37*(8), 1493-1506.

Schaeffer, B., Eggleston, V. H., & Scott, J. L. (1974). Number development in young children. *Cognitive Psychology*, *6*(3), 357-379.

Schneider, R. M., Sullivan, J., Marušič, F., Biswas, P., Mišmaš, P., Plesničar, V., & Barner, D. (2020). Do children use language structure to discover the recursive rules of counting? *Cognitive Psychology*, *117*, 101263.

Sella, F., Slusser, E., Odic, D., & Krajcsi, A. (2021). The emergence of children's natural number concepts: Current theoretical challenges. *Child Development Perspectives.* DOI: 10.1111/cdep.12428

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research, 13*(4), 251-271.

Shusterman, A., Cheung, P., Taggart, J., Bass, I., Berkowitz, T., Leonard, J. A., & Schwartz, A. (2017). Conceptual correlates of counting: Children's spontaneous matching and tracking of large sets reflects their knowledge of the cardinal principle. *Journal of Numerical Cognition, 3*(1), 1-30.

Shusterman, A., Slusser, E., Halberda, J., & Odic, D. (2016). Acquisition of the cardinal principle coincides with improvement in approximate number system acuity in preschoolers. *PloS one, 11*(4), e0153072.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257-268.

Slusser, E., Ditta, A., & Sarnecka, B. (2013). Connecting numbers to discrete quantification: A step in the child's construction of integer concepts. *Cognition, 129*(1), 31-41.

Spaepen, E., Gunderson, E. A., Gibson, D., Goldin-Meadow, S., & Levine, S. C. (2018). Meaning before order: Cardinal principle knowledge predicts improvement in understanding the successor principle and exact ordering. *Cognition, 180*, 59-81.

Stevenson, M., & Sergeant, E. (2021). Package 'epiR'.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics* (Fourth S., editor) New York.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, *37*(5), 360-363.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274-290.

Wagner, K., Chu, J., & Barner, D. (2019). Do children's number words begin noisy?. *Developmental science*, *22*(1), e12752.

Wagner, J. B., & Johnson, S. C. (2011). An association between understanding cardinality and analog magnitude representations in preschoolers. *Cognition*, *119*(1), 10-22.

Wagner, K., Kimura, K., Cheung, P., & Barner, D. (2015). Why is number word learning hard? Evidence from bilingual learners. *Cognitive Psychology*, *83*, 1-21.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155-193.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, *24*(2), 220-251.

Yantz, C. L., & McCaffrey, R. J. (2009). Effects of parental presence and child characteristics on children's neuropsychological test performance: third party observer effect confirmed. *The Clinical Neuropsychologist*, *23*(1), 118-132.

CHAPTER 3

The Development of Subitizing in Bilingual Children

Elisabeth Marchand, Nina Schoener, Jocelyn Sandoval-Franquez, Kelly Kendro and David Barner[1]

[1]Department of Psychology, University of California, San Diego

Abstract

What role does language-specific experience play in the development of numerical knowledge? Previous studies have probed this question by testing bilingual learners, who have different linguistic representations of number across their two languages. Here, we take this approach to study the effect of language-specific experience on subitizing, a measure of children's ability to rapidly estimate small sets. Specifically, we tested 66 Spanish-English and German-English bilinguals, aged 3 to 6, and found that bilinguals made more accurate verbal estimates of small sets in their dominant number language (the language in which they could count the highest). This was despite the fact that all children in the study were able to accurately count much larger sets, and were classified as "cardinal principle knowers" in both of their two languages. These results provide evidence for early emerging individual differences in estimation abilities that are due to linguistic experience. Also, they suggest that even after children have begun to learn more advanced skills (e.g., accurate counting of large sets), individual differences in earlier learned skills (e.g., subitizing of small sets) may persist. Different language specific experiences may therefore impose persistent effects on numerical development, with implications not only for bilingual learners, but also for monolinguals with impoverished exposure to number language.

Introduction

Although many animals have the capacity to represent approximate numerical magnitudes, only humans have symbols for large exact numbers, a fact that is often attributed to our species-specific capacity for natural language (Carey & Barner, 2019; Gordon, 2004; Le Corre & Carey, 2007; Pica et al., 2004; Spaepen et al., 2013; Spelke, 2017). However, the precise role that language plays in the acquisition of exact number remains unclear, in part because linguistic and non-linguistic capacities emerge and change together during development, making their respective roles difficult to disentangle. In efforts to isolate the role of language from non-linguistic factors, recent studies have focused increasingly on cross-linguistic comparison, as well as on studies of bilingual learners, who make it possible to distinguish between language-specific knowledge and knowledge that is not specific to a particular language. In the present study we adopted this logic, and investigated language-specific processes involved in subitization, a mechanism that has been argued to play a fundamental role in building exact representations of numbers.

Previous studies of bilinguals have played an important role in revealing how some forms of numerical abilities rely on language-specific experiences. For example, in one important study, Spelke and Tviskin (2001) trained English-Russian bilinguals in each of their two languages on exact and approximate calculations (e.g., additions in base 6, cube root estimation). They found that when bilinguals were trained on problems in one language, they were slower and less accurate when asked to solve similar problems in their second language, but only when the calculations involved exact numbers. In contrast, approximate calculations were solved equally well in their two languages. Similar findings have been observed with other forms of exact arithmetic calculations (e.g., multiplication and subtraction) and in bilingual speakers of different languages

(Saalbach et al., 2013; Van Rinsveld et al., 2015; Venkatraman et al., 2006). Behavioral findings have also been corroborated by neuroimaging techniques which show that exact computations, unlike approximate calculations, systematically recruit language areas (Grabner et al., 2012; Lin et al., 2012; Mondt et al., 2011; Salillas & Wicha, 2012). Overall, these findings suggest that numerical skills that draw on stored math facts and exact computations are not only language-dependent but also language-specific, and don't readily transfer across languages.

While previous studies have found evidence for the role of language-specific experiences in arithmetic, less is known about how bilinguals represent other forms of numerical knowledge, and how these representations emerge in development. For example, considerably less is known about earlier learning processes in bilinguals, including how initial exact meanings are learned, and how these early linguistic expressions of exact number become associated with approximate number representations. In particular, little is known about whether these forms of knowledge transfer automatically across languages, or if learning in a child's two languages is independent. In one recent study of this question, Marchand et al. (2020) investigated the estimation abilities of 5- to 7-year-old bilingual children. Marchand et al. found that bilinguals estimated arrays differently across their two languages — making more accurate estimates in their more dominant language than in their less dominant one (as determined by how high they could count in each language). Although many of these children could count higher in one language than in the other, the reported difference in estimation accuracy was found even for numbers that children could produce in both languages, indicating that estimation differences weren't simply due to lacking words in one of the two languages.

These findings regarding estimation in bilinguals support two conclusions. First, they show that changes in estimation skills in early childhood do not depend purely on the development of

non-verbal number representations. Previous studies of monolingual children have debated whether individual differences in the precision of non-verbal approximate number representations are related to both later arithmetic abilities (Sullivan et al., 2016) or to verbal estimation abilities (Guillaume & Gever, 2016). However, this literature has focused less on how approximate number representations become related to language, and whether differences in linguistic knowledge of counting structures impacts estimation abilities. Thus, the data from Marchand et al. provide important evidence that developmental changes in estimation ability depend not only on the maturation of non-verbal number, but also on individual differences in children's linguistic number representations. Second, these results show that linguistically-mediated disparities in numerical knowledge — which presumably can explain individual differences between monolinguals too — begin to emerge relatively early in development, by at least 5 years of age.

Critically, however, the results of Marchand et al. leave open when this linguistically-mediated disparity in estimation ability first begins to emerge. This is important because some theories have argued that children begin number word learning by associating individual words like "one", "two", and "three" with cardinalities (Carey, 2004). However, Marchand et al. mainly tested larger numbers (i.e., 8 and higher), and the youngest children were 5 years of age. Therefore, in the current study we investigated the earliest moments of number word acquisition in bilingual children to assess when disparities in estimation abilities might first emerge in development, focusing on numbers within the subitizing range of 1-3. In this way, we asked when language-specific experiences begin to impose different learning trajectories on number knowledge.

Previous studies find that the process of associating number words to cardinalities begins by at least the age of 2 (Fuson et al., 1982; Gelman & Gallistel, 1978; Wynn, 1990, 1992), though relatively little is known about the relative roles of linguistic and non-linguistic sources of change.

Initially, many children go through a phase during which they are able to recite part of the count list (e.g., number words *one* through *five*) without understanding the meanings of individual words, similar to their rote knowledge of songs, or the alphabet. However, not long after, children begin to build associate mappings between small number words and cardinalities. Using Wynn's (1990, 1992) Give-a-Number task (Give-N), studies in this literature have found that children begin by associating the word *one* with sets containing exactly one object: when asked to give *one* object, they correctly provide one object, but can't reliably give correct amounts when asked to give larger numbers. For this reason, children at this stage are often called "one-knowers". Some months later, children learn the meaning of *two*, and respond correctly when asked to give *one* or *two* objects, but not larger numbers. These children are called "two-knowers." Following the same pattern, children become "three-knowers" and sometimes "four-knowers" over a period of many months. Finally, after moving through these "subset" stages, children seem to realize that they can use the counting procedure to generate sets of different sizes, and use the counting routine that they learned much earlier to accurately give sets larger than 3-4. Because children at this stage seem to understand that counting relates to cardinalities, they are called Cardinal Principle or Counting Principle knowers (CP-knowers; for discussion, see Cheung et al., 2017; Davidson et al., 2012; Schneider et al., 2020; Sella & Lucangeli, 2020). This general pattern of development has been found in many studies across many different languages (Almoammer et al., 2013; Barner et al., 2009; Ceylan & Aslan, 2018; Condry & Spelke, 2008; Davidson et al., 2012; Jara-Ettinger et al., 2017; Le Corre & Carey, 2007; Le Corre et al., 2006, 2016; Li et al., 2003; Marchand & Barner, 2019; Meyer et al., 2020; Negen & Sarnecka, 2012; Nikoloska, 2009; Piantadosi et al., 2014; Sarnecka & Carey, 2008; Sarnecka et al., 2007; Sarnecka & Lee, 2009; Sarnecka et al., 2018; Schneider et al., 2020; Slusser et al., 2013; Spaepen et al., 2018; Wagner et al., 2015; Wynn, 1990,

1992). While this process of associating number words with cardinalities has been well documented in monolingual children, less is known about this early learning process in bilingual learners.

In one study of this question, Wagner et al. (2015) tested French-English and Spanish-English bilingual children in their two languages on the Give-N task and found that bilinguals had different knower levels in their first (L1) and second (L2) languages. In particular, they found that although children classified as CP-knowers in one language were generally CP-knowers in their other language too, a different pattern was found for children who could not yet accurately count. In particular, children who were identified as 1-, 2- or 3-knowersin one language had a different knower level in their second language about 2/3 of the time. For example, most children who were classified as, e.g., 2-knowers in the L1 were not classified as 2-knowers in their L2, despite being able to count well past *two* in both of their two languages. Similarly, a more recent study by Sarnecka et al. (2021) compared the early number knowledge of Spanish-English bilinguals across their two languages and found that children performed significantly better in English (their language of instruction) across multiple tasks (i.e., counting to 10, counting 6 objects and the Give-N task). Together, these studies suggest that knowledge of small number words that is acquired in one language does not automatically transfer to a child's second language. Although bilingual children exhibit different knowledge of small number words prior to learning to count, as shown by Wagner et al. this difference between languages appears to resolve itself when children learn to count, and become CP-knowers. As described above, almost all children in their study who were CP-knowers in one language were also accurate counters in their second language, too. Taken in isolation, these results might be interpreted as evidence that once bilingual children become CP-knowers their knowledge of small number words across their two languages equalizes. Compatible

with this logic, to be classified as a CP-knower in the Give-N task, children must typically provide accurate responses not only to larger numbers, but also to small ones. However, an alternative hypothesis is that, while bilingual children may have reached the same minimum standard to be classified as CP-knowers in both languages, they may still exhibit language-specific differences in how robustly they represent small number words, in ways not detectable by the Give-N task. Critically, to be credited with knowledge of a number word like "two" on the Give-N task, a child need only respond correctly on 2/3 of requests to give this number. Consequently, a child who has very strong associative mappings between small number words and cardinalities in one language but relatively weak associations in their second language might still be credited with the same knowledge of these words in their two languages, so long as their associations are strong enough to respond correctly on 2/3 trials in each case. Further, given the finding that counting abilities transfer across a child's two languages, it's also possible that bilinguals respond successfully to requests for smaller numbers in their weaker language without drawing on associative mappings at all, and simply count when asked to give smaller sets, thereby masking the differences in knowledge across their two languages.

One way to address this issue is to assess bilingual children's knowledge of small number words in a task that does not permit counting, via tests of "subitizing" (Kaufman et al., 1949; Starkey & Cooper, 1995). Typically, in a subitizing task, participants are presented with flashed arrays of dots (1 to 4 dots) and are asked to make verbal numerical estimates of the cardinalities. Because of the rapid presentation of stimuli, the subitizing task discourages counting such that participants must rely instead on associations between words and sets. A subitizing task might also be more sensitive to small differences in knowledge than Give-N because it can be used to measure not only accuracy on individual trials, but also differences in response time. Finally, beyond the

benefit of having a stronger measure of the mappings between number words and cardinality, studying knowledge of subitizing skills across languages in bilinguals is on its own important because subitizing performance has been shown to be correlated with non-symbolic arithmetic, number line estimation, and counting (LeFevre et al., 2010, 2022) and has been argued to play an important role in the acquisition of number words (Carey, 2004).

In the current study, we tested bilingual CP-knowers' subitizing abilities across their two languages to investigate the role of language-specific experiences in the acquisition of number words. To assess generalizability across different language and SES groups, we tested two samples of bilingual children: Spanish-English children from low SES families, and German-English bilinguals from high SES families. If bilinguals, in the process of becoming CP-knowers, build equally robust mappings between number words and their cardinality across languages, then we should not find differences in subitizing abilities across languages. However, if these mappings still remain reliant upon language-specific experiences across languages even after children become CP-knowers, then we should see differences in subitizing performances across languages.

Method

Participants

We tested a total of 113 bilingual children. Forty-seven children were excluded from analysis because of (1) being a subset-knower (n = 20), (2) reaching the same Highest Count in each language (n = 5), (3) language delay (n = 3), (4) failure to complete all 6 tasks (n = 8) and (5) experimenter error, parental interference, or technological issues (n = 11). Our final sample included 66 children (36 Spanish-English and 30 German-English bilinguals), aged 3.47 to 5.98 years (M = 4.98 years). Additionally, we tested 73 English monolingual children for comparison. Thirty-one children were excluded from analyses because of (1) being a non-knower or subset-

knower (n = 17), (2) failure to complete all 3 tasks (n = 6), or (3) experimenter error, parental interference, or technological issues (n = 8). Our final sample of monolingual children included 42 participants, aged 2.75 to 5.87 (M = 4.50 years). All monolingual and bilingual participants were recruited from either a participant database, via communications with preschools, or using online recruitment tools such as Facebook and Children Helping Science. All participants lived in the United States with most living in California. The majority of German-English bilingual participants were recruited from private German-speaking daycare centers or schools; though we did not collect detailed socioeconomic information, most of these participants were living in affluent areas in California (e.g., Silicon Valley). The majority of Spanish-English bilingual participants were recruited both through word-of-mouth and through primarily hispanic public school districts in California; these districts had average median household incomes below the average for California according to the 2021 Census Reporter (U.S. Census Bureau, 2021). Recruitment was performed by either German-English or Spanish-English bilingual research assistants. As part of the sign-up and consent process, parents received a short questionnaire with questions about the child's first language, second language, and whether the child was bilingual. Children who were not identified as bilingual by caregivers were excluded from the bilingual sample and took part in the study as monolinguals.

Materials and procedure

Because this study took place during the COVID-19 pandemic, all participants were tested online via Zoom. Bilingual children received two blocks of tasks—an English block and a Spanish or German block—the order of which varied between participants. Within each block, children were always presented with 3 tasks in the same order: (1) Give-a-Number task, (2) Fast Cards, and (3) Highest Count task. Both groups of bilinguals (Spanish-English and German-English) were

tested by a native speaker of either German or Spanish who also spoke English. The testing session lasted approximately 30 minutes. The procedure for monolingual children was identical to that of bilinguals except that they were administered only the English block of tasks and their testing session lasted approximately 15 minutes. Monolingual children were tested by a native English speaker. All children received a small prize for their participation at the end of the testing session.

Give-a-Number Task

This task was adapted from Wynn (1990) and its goal was to determine whether children were CP-knowers or not, since only CP-knowers were included in our main analyses. Before the testing session, parents were asked to gather a container (usually a plate) and 10 small identical objects (e.g., almonds, coins, raisins, etc.). For bilinguals, the task was administered in both of the child's two languages and the experimenter began the task by addressing the child in the language to be assessed first (e.g., English, German, or Spanish). The experimenter first asked the child to put a certain number of objects into the container (e.g., "Can you put three almonds on the plate?"). After this first prompt, children were asked to count to verify that they had provided the right number, and if they chose to fix their answers, only final responses were recorded (Gibson et al., 2019). The trial structure followed a titration procedure. Children were always asked for one first, and then two, and if they succeeded on both trials, the experimenter then asked for three. Otherwise, they asked for one. The subsequent requests depended on the child's pattern of response: if the child succeeded on a trial requesting N, the experimenter asked for N+1, but if the child failed, they asked for N-1. The highest request was six. Participants were credited as CP-knowers if they correctly gave N objects at least 67% of the time when asked for N (and not other requests) for numbers up to 6. Bilingual children were tested by the same experimenter in both

languages. Only children classified as CP-knowers in both languages were included in the main analyses.

Fast Cards Task

This task was adapted from Le Corre and Carey (2007) and was used to assess children's subitizing abilities. Participants were presented with dot arrays ranging in magnitude from 1 to 8. Critical trials tested sets of 1 to 4 and additional trials tested 6 and 8. The larger sets were included in exploratory analyses to assess children's mapping to large numbers. For each language that a child spoke (e.g., English and Spanish), all magnitudes were presented 4 times each, for a total of 24 trials. The trials were divided into 4 blocks containing one of each of the 6 magnitudes. Within each block, the order of trials was pseudo-randomized such that it differed across blocks. The order of blocks was then pseudo-randomized to create 2 orders. For all participants, half of the trials controlled for total surface area while the other half controlled for diameter. In the trials controlling for total surface area, the total surface of all the dots (i.e., sum of all diameters) was kept constant while in the trials controlling for diameter, the diameter of dots was kept constant. In addition to the 2 orders of blocks, we created 2 fixed orders for the controlled parameters such that for half of the participants, the first trial controlled for diameter (and then alternated with total surface area) while for the other half of participants, the first trial controlled for total surface area. Half of the children were tested in English first, and for the other half, Spanish/German was assessed first. Hence, participants were assigned to one of 16 conditions in a 2 (English vs Spanish/German first) x 2 (Order 1 or 2 of blocks) x 2 (Order 1 or 2 of controlled parameters) x 2 (language within-subject) design. Dots were displayed on the screen for 1 second. All dots were white and displayed on a black background.

The instructions were provided to the child in the language dictated by the condition in which she was in (Spanish/German or English depending on language at test) or in English for monolingual participants. Participants were instructed to look at the dots and guess how many there were in the following way: "*Let's play a guessing game! Some dots are going to flash on the screen for just a second. I want you to guess how many dots there are! Ready? Great! How many dots did you see on the screen?"*. Two practice trials were modeled by the experimenter (with sets of 5 and 7) to ensure that children understood that they needed to guess without counting. Noninformative verbal encouragement was given to the children to keep them motivated, regardless of response's accuracy (e.g., "That's a good guess!").

Highest Count Task

This task had two goals. First, following the logic of Wagner et al. (2015), we used it to identify bilingual children's dominant number language. Second, we used the task as a general proxy for counting experience. Children were asked to count a set of 30 cartoon cats presented on a Powerpoint slide, in 4 rows of about 8 cats. Children who reached 30 were prompted to count further, which in the case of bilingual children served to reduce the possibility of reaching the same Highest Count in both languages. Bilingual children's dominant number language was defined as the language in which they could count highest before stopping or making an error. Below we refer to the dominant number language (NL) as their NL1, and their non-dominant number language as their NL2.

Results

For all analyses, we used mixed effects logistic regression models constructed in R using the lme4 package version 1.1-27.1 (Bates et al., 2015). All data and model outputs are available

here: cite. All planned analyses were preregistered and were conducted as planned unless otherwise stated.

Highest Count

We first compared bilingual children's highest counts across their two languages to identify their dominant Number Language (NL1; see Figure 1). Their non-dominant language was labeled their NL2. Among the Spanish-English bilingual children, 20 counted higher in English and therefore their NL1 was identified as English. The NL1 of the remaining 16 Spanish-English children was Spanish. Among the German-English bilinguals, English was the NL1 for 18 children, and German was the NL1 for the remaining 12. On average, both Spanish-English and German-English bilinguals counted slightly higher in English than in their other language (for Spanish-English: M_{English} = 20.8, *range* = 5 to 100; M_{Spanish} = 15.8, *range* = 5 to 59; for German-English: M_{English} = 22.3, *range* = 5 to 35; M_{German} = 18.8, *range* = 11 to 32). Children had a mean highest count of 24.2 in their NL1 (*range* = 6 to 100) and 14.4 in their NL2 (*range* = 5 to 59), which was significantly different (V = 2211, $p < .0001$).

Children in our monolingual sample had highest counts similar to the NL1 counts of our bilingual children (*range* = 3 to 31; M = 20.5) and the difference between these two datasets was not statistically significant (p = .07). However, the highest count of monolinguals did differ significantly from the NL2 highest counts of bilinguals ($t(90) = -3.80$, $p < .001$).

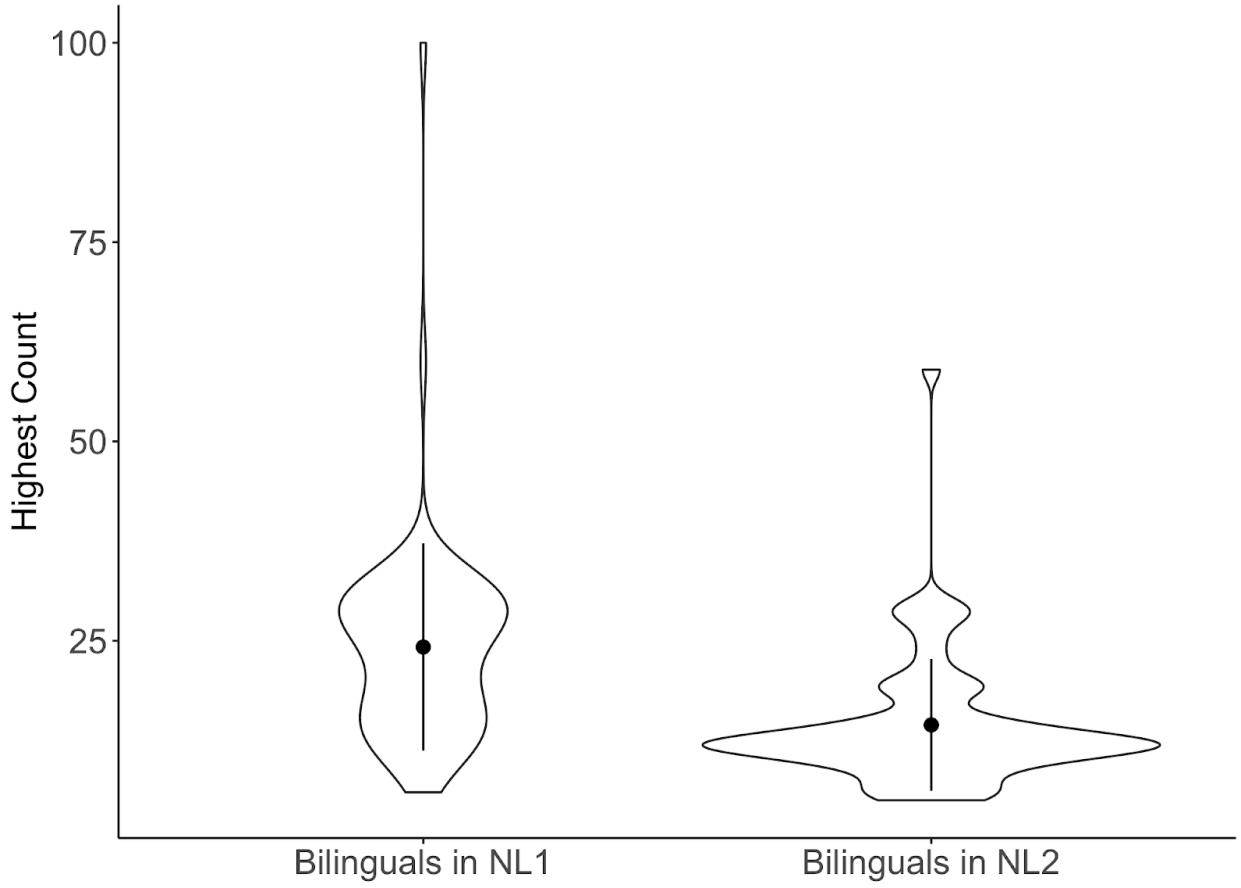

Figure 3.1: Bilingual participants' Highest Counts by dominant language.

Note. NL1 stands for children's dominant language and NL2 is their less dominant language. There were 38 English-dominant, 12 German-dominant and 16 Spanish-dominant children.

Subitizing performance of bilinguals across languages

Before conducting our main analyses comparing the subitizing abilities of bilingual children across their two languages, we excluded all responses that were not provided in the format of a unique verbal estimate (n = 50/3168), e.g., "too many" or "poquitos". Responses that were 10 times larger than the mean estimate for each numerosity presented were also excluded, as pre-registered (n = 17/3168). All analyses were performed on the remaining 3101 responses.

Next, we conducted preliminary analyses to test whether there were any effects of Task Order or Bilingual Type (i.e., German-English vs. Spanish-English bilinguals). To do this, in separate analyses, we predicted Accuracy of responses (coded as 0 or 1) from Task Order and Bilingual Type, and in both cases, Participant was coded as a random factor. Since we found no effect of Task Order or Bilingual Type (all *p*s > .10), these factors were excluded from subsequent analyses and all bilingual children were analyzed together.

Main Analyses

In our main analyses, we were interested in whether bilingual CP-knowers subitized differently across their two languages (see Figure 2). We analyzed responses for numerosities 1 to 4 using model comparison. In our first model, we predicted Accuracy (coded as 0/1) from Age (in years) and Numerosity (1 to 4) with Participant coded as a random factor. In this model, both Age (*Estimate* = 1.32, *SE* = .30, *Z* = 4.43, $p < .0001$) and Numerosity (*Estimate* = -1.62, *SE* = .12, *Z* = -13.29, $p < .0001$) were significant. As expected, Accuracy improved as a function of Age and decreased as the Numerosity of sets increased. In our second model, we added the effect of Number Language (i.e., NL1 vs. NL2) to a model that otherwise was identical to the first model. In this second model, Age (*Estimate* = 1.34, *SE* = .30, *Z* = 4.43, $p < .0001$) and Numerosity (*Estimate* = -1.63, *SE* = .12, *Z* = -13.30, $p < .0001$) remained significant. Importantly, we also found that Number Language was significant (*Estimate* = -.57; *SE* = .18, *Z* = -3.16; *p* = .002). An ANOVA comparing these models found that adding Number Language significantly contributed to the explanatory power of the model ($p < .01$), suggesting that bilingual CP-knowers subitize differently across their two languages. In particular, as can be seen in Figure 2, bilingual children subitized more accurately in their NL1 relative to their NL2. Furthermore, an exploratory analysis

found that these effects persisted when Highest Count was added to models ($p < .05$)[16], suggesting that the effect of language was not due to different amounts of counting knowledge. Although this effect was small, it provides evidence that bilingual children who are classified as CP-knowers in both languages still have reliably different mappings between number words and cardinalities across languages.

Next, following our pre-registered plan, we ran a third model in which we added the interaction between Number Language and Numerosity to our second model. In this third model, while Age, Numerosity, and Number Language remained significant (all ps < .05), the interaction was not ($p = .07$) and an ANOVA comparing this model to the second model revealed that this last model did not contribute significantly to the explanatory power of the model ($p = .06$). This suggests that the differences in subitizing' accuracy between languages was somewhat uniform across numerosities, and that if differences existed they were relatively small.

[16] Here, we deviate from the pre-registered analyses as they were not the best fit for the data we collected. In the model reported here, we predicted Accuracy from Age, Numerosity (1-4), Number Language, and the child's Highest Count in their NL1 and in their NL2. As reported, all factors were significant (all ps < .05). The values of Highest Count in NL1 and in NL2 were also significantly correlated ($r_\tau = .68$, $p < .0001$).

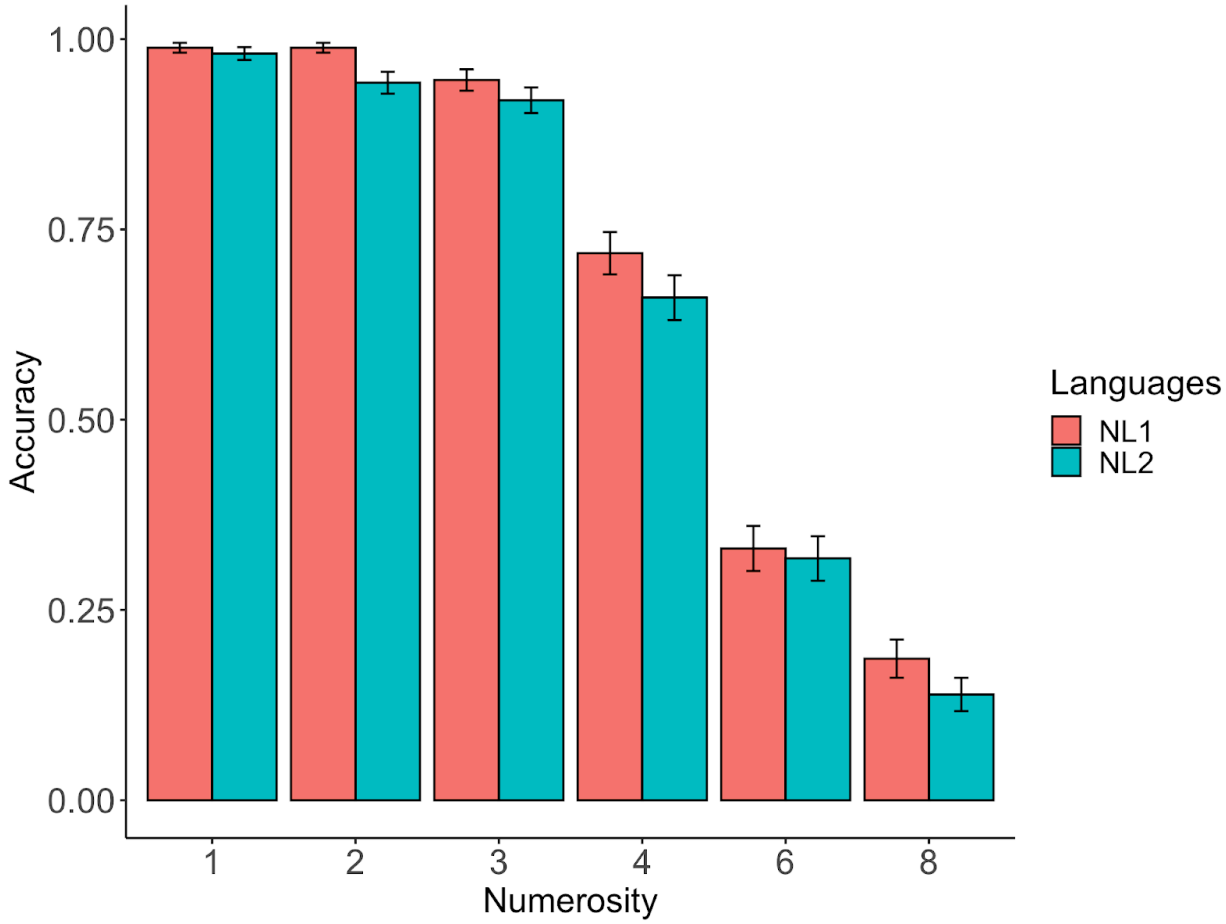

Figure 3.2: Subitizing Accuracy for each Numerosity across Bilinguals'two languages.

Note. The y axis represents children's subitizing accuracy out of 4 trials for each numerosity. Numerosity is displayed on the x axis. NL1 and NL2 respectively represent children's most dominant Number Language and less dominant Number Language. Bars represent standard errors.

Exploratory Analyses

We next conducted a series of exploratory analyses to address: 1) whether the linguistically mediated differences in subitizing observed in bilingual children extend to larger numbers (i.e., 6 and 8) and, 2) how the subitizing skills of monolingual children compared to those of bilingual children across their two languages.

Estimation of larger numbers

In our first set of exploratory analyses, we asked whether bilinguals' estimates for numbers outside the subitizing range differed across their two languages. To do this, we analyzed only responses provided for arrays of 6 and 8 dots, and used the same model comparison approach described in our main analyses section. In a model that predicted Accuracy from Age and Numerosity (and Participant as a random factor), both factors were significant (Age: *Estimate* = .59; *SE* = .18; *Z* = 3.27; $p < .01$; and Numerosity: *Estimate* = -0.58; *SE* = .09, *Z* = -6.65; $p < .0001$). In a second model, we added the main effect of Number Language, but this factor was not significant, and the model did not improve fit to the data (both *p*s > .05). Similarly, the interaction between Numerosity and Number Language (model 3) did not add explanatory power to the first model ($p > .05$). This suggests that Accuracy improved with Age and decreased with Numerosity similarly across languages. However, possibly explaining why no difference was found, as shown in Figure 2 the average accuracy for sets of 6 and 8 was overall very low (sets of 6: 33% in NL1 and 32% in NL2; for sets of 8: 19% in NL1 and 14% in NL2). We return to this issue in the Discussion section.

Comparing monolinguals and bilinguals in their NL1 and NL2

In our final exploratory analysis, we compared bilinguals' subitizing skills to those of monolingual CP-knowers. We did this for two reasons: first to assess whether bilingual children show an overall delay in subitizing compared to monolingual children and second, to assess how the mappings in bilinguals' NL1 and NL2 compared to those of monolingual children who had the same knower level.

We first excluded from the monolingual dataset all responses that were either not in the format of a unique verbal estimate (n = 42/1008) or that were 10 times larger than the mean estimate for each numerosity (n = 1/1008). The remaining 965 responses were compared to those

of bilingual children in their NL1 and NL2. As we did for the main analyses, we restricted the dataset to responses for sets within the subitizing range of 1 to 4 dots.

We compared monolinguals and bilinguals' responses in NL1 and NL2 in two different sets of model comparisons[17]. Table 1 shows the average response and Accuracy of monolingual children. To compare monolinguals and bilinguals' responses in their NL1, we first predicted Accuracy from Age and Numerosity (Participant was also coded as a random factor) and found that both factors were significant (Age: *Estimate* = 1.37; *SE* = .29, *Z* = 4.78; $p < .0001$; Numerosity: *Estimate* = -2.22; *SE* = .20, *Z* = -11.37; $p < .0001$). In our second model, we added the main effect of bilingual status (either bilingual or monolingual) and found no significant effect of this factor ($p = .36$). This suggests that in our sample, the subitizing abilities of monolinguals and bilingual children in their NL1 did not differ. Next, to compare monolinguals and bilinguals' NL2 responses, we predicted Accuracy from Age (*Estimate* = 1.45; *SE* = .31, *Z* = 4.68; $p < .0001$) and Numerosity (*Estimate* = -1.85; *SE* = .15, *Z* = -12.32; $p < .0001$) and in a second model, we added the effect of bilingual status. In this second model, all factors were significant (Age: *Estimate* = 1.63; *SE* = .32, *Z* = 5.11; $p < .0001$; Numerosity: *Estimate* = -1.85; *SE* = .15, *Z* = -12.32; $p < .0001$; Bilingual Status: *Estimate* = 1.05; *SE* = .47, *Z* = 2.22; $p = .026$). Because the main effect was significant, we added in a third model the interaction between Bilingual status and Numerosity, and this factor was also significant (*Estimate* = -.78; *SE* = .34, *Z* = -2.28; $p = .023$) as were all others (Age: *Estimate* = 1.65, *SE* = .32, *Z* = 5.17, $p < .0001$; Numerosity: *Estimate* = -1.63, *SE* = .17, *Z* = -9.82, $p < .0001$; Bilingual status: *Estimate* = 3.84, *SE* = 1.34, *Z* = 2.88, $p < .01$). Taken together, these

[17] As indicated in the pre-registration, we originally planned to obtain a dataset of monolingual subset-knowers to which we could compare the subitizing skills of bilingual children in their NL2. However, due to recruitment difficulty during the pandemic and the large sample size needed, we focused our efforts on recruiting monolingual CP-knowers only.

findings suggest that bilinguals who have comparable NL1 subitizing abilities to monolingual children nevertheless differ from monolinguals when considering their NL2.

Table 11: Average estimate and accuracy of response per Numerosity and Number Language in the Fast Cards Task for Monolingual children.

Numerosity	Responses		Accuracy	
	M	*SD*	*M*	*SD*
1	1.07	0.60	0.99	0.11
2	2.00	0.00	1.00	0.00
3	3.22	1.51	0.92	0.28
4	4.29	1.86	0.67	0.47

Note. M represents the mean and *SD* the standard deviation.

Discussion

In this study, we investigated language-specific differences in children's representations of small number words by testing subitizing abilities of bilingual CP-knowers across their two languages. Specifically, we presented Spanish-English and German-English CP-knowers with flashed arrays of small sets (1 to 4) and asked them to estimate these sets in each of their two languages. We found that bilingual children subitize differently across their two languages and specifically, that they are more accurate in the language in which they have the most experience counting. These findings provide evidence that despite meeting the criteria to be classified as CP-knowers in both languages, bilingual children still have different mappings between number words and small cardinalities across their two languages. This suggests that language-specific experience plays an important role in the development of mappings between number words and cardinalities across languages, and that these differences persist even after children have progressed to later stages of development.

The results of this study allow us to integrate previous findings regarding number word learning and numerical estimation in bilingual children. In particular, our results suggest that the disparity in knowledge of number word meanings found in bilingual subset-knowers (Wagner et

al., 2015) persists even after children learn to accurately count large sets and become CP-knowers. This is compatible with previous findings that older bilingual children show differences in their estimation skills across languages (Marchand et al., 2020) and suggests that these differences begin to emerge from the early moments of number word learning. It is debated in the literature whether the processes of subitizing and the estimation of large numbers rely on the same underlying mechanisms; some argue that subitizing taps into a Parallel Individuation system - a system that keeps track of ~ 4 individual items in parallel in working memory - while the second relies on the Approximate Number System (or Analog Magnitude System) - a system that allows us to apprehend approximately the magnitude of large sets (Carey, 2004; Feigenson et al., 2004). While our results can't directly address this issue, they provide evidence that language-specific experience impacts both small and large numbers and that language-related disparities emerge at an early age in bilingual children.

This finding has implications for our understanding of number word learning in bilinguals, but is also relevant to theoretical debates regarding monolingual learners. Bilinguals' successful application of the counting procedure in both languages, despite having different mappings across languages, raises questions about the causal relationship between learning the meanings of small number words and becoming a CP-knower. According to bootstrapping accounts (e.g., Carey, 2004, 2009), learning the meanings of small number words is an essential prerequisite to becoming a CP-knower. In particular, on this account children notice that each time they add one object to a set, they should count up one word in the count list - a relation, which, when generalized to all possible numbers, should allow them to accurately count and build indefinitely large sets. An alternative, however, is that learning to count is relatively independent of mapping small number words to their meanings, and that these are two distinct, though parallel processes (Barner, 2017).

On this account, CP-knowers initially perform the Give-N task using rote procedures, and only gradually learn, through experience applying this procedure, how counting is related to iterative processes of addition. Although our data are hardly definitive, they are compatible with the idea that children can become CP-knowers given relatively different degrees of knowledge of smaller number words, as argued by the "parallel process" model of number word learning. Our study also suggests that there may be individual differences in the strength of small number word knowledge between monolingual children who classify as CP-knowers, and that possibly, some children may skip some subset-stages altogether. Future studies should further explore this possibility.

Our set of exploratory analyses uncovered two results: 1) the cross-linguistic differences found for small numerosities did not extend to the two larger sets (i.e., 6 & 8) assessed in our task and 2) bilingual children's subitizing skills in their NL1 were similar to the subitizing skills of monolinguals, unlike bilinguals' subitizing skills in their NL2. With respect to the estimation of larger sets, the lack of a significant difference in accuracy is perhaps surprising given that older bilinguals show differences in estimation across languages (Marchand et al., 2020). However, one possible explanation for this discrepancy is that, because children in the current study were much younger than those in Marchand et al. (2020), they had not yet begun to form strong associative mappings for larger numbers in either language, and therefore failed to show linguistically mediated differences due to floor effects. As evidence for this, children' accuracy for sets of 6 and 8 was low, which contrasts with Marchand et al. (2020), in which the performance for these types of sets was much more accurate than for larger sets such as 60, 80, 98. With respect to the comparison of monolingual and bilingual children, our results suggest that bilinguals are neither delayed nor advanced relative to monolingual children. When tested in their dominant number language bilinguals resembled monolingual English-speaking children, whereas they performed

slightly worse when tested in their non-dominant language. This result corroborates findings from other studies with bilingual children that did not find differences between monolingual English speakers and Spanish-English children tested in English (i.e., bilinguals who attended preschool programs in English; Sarnecka et al., 2021).

In sum, the current study shows that subitizing skills differ across bilingual children's two languages, such that they have stronger mappings in the language in which they can count the highest, despite being able to count beyond the small number range in both languages. These results have potentially important implications for mathematics education. In particular, researchers should not assume that bilinguals' numerical skills are identical across their two languages. Children may reach a proficient level in math in their language of instruction, but may nevertheless be unable to apply this knowledge in their heritage language without additional language-specific training. Future studies should investigate optimal methods for training these mappings in bilinguals, and whether it is better to explicitly teach children to translate across languages, or instead offer separate language-specific training of the mappings between small sets and their cardinalities.

Acknowledgements

Chapter 3, in full, is currently being prepared for submission for publication of the material. Marchand, E., Schoener, N., Sandoval-Franquez, J., Kendro, K., and Barner, D. The dissertation author was the primary investigator and author of this paper.

This work was supported by the Social Sciences and Humanities Research Council of Canada via a fellowship to E.M., a James S. McDonnell Foundation award to D.B., and a National Science Foundation award (ID: 2000827) to D.B. We would like to thank all the children and families who participated in this project. Special thanks are also due to Deyanira Camarena and

Angie Rodriguez-Verdin for their help in collecting data. Finally, we thank the members of the Language and Development Lab at UCSD for their helpful feedback on this work.

References

Almoammer, A., Sullivan, J., Donlan, C., Marušič, F., O'Donnell, T., & Barner, D. (2013). Grammatical morphology as a source of early number word meanings. *Proceedings of the National Academy of Sciences*, *110*(46), 18448-18453.

Barner, D. (2017). Language, procedures, and the non-perceptual origin of number word meanings. *Journal of child language*, *44*(3), 553-590.

Barner, D., Libenson, A., Cheung, P., & Takasaki, M. (2009). Cross-linguistic relations between quantifiers and numerals in language acquisition: Evidence from Japanese. *Journal of Experimental Child Psychology, 103*(4), 421-440.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01.

Carey, S. (2004). Bootstrapping & the origin of concepts. Daedalus, 133(1), 59-68.

Carey, S. (2009). Where our number concepts come from. The Journal of philosophy, 106(4), 220.

Carey, S., & Barner, D. (2019). Ontogenetic origins of human integer representations. *Trends in cognitive sciences*, *23*(10), 823-835.

Ceylan, M., & Aslan, D. (2018). Cardinal Number Acquisition of Turkish Children. *Journal of Education and e-Learning Research*, *5*(4), 217-224.

Cheung, P., Rubenson, M., & Barner, D. (2017). To infinity and beyond: Children generalize the successor function to all possible numbers years after learning to count. *Cognitive psychology*, *92*, 22-36.

Condry, K. F., & Spelke, E. S. (2008). The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology: General*, *137*(1), 22.

Davidson, K., Eng, K., & Barner, D. (2012). Does learning to count involve a semantic induction?. *Cognition*, *123*(1), 162-173.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*(7), 307-314.

Fuson, K. C., Richards, J., & Briars, D. J. (1982). The acquisition and elaboration of the number word sequence. In *Children's logical and mathematical cognition* (pp. 33-92). Springer, New York, NY.

Geary, D. C., vanMarle, K., Chu, F. W., Rouder, J., Hoard, M. K., & Nugent, L. (2018). Early conceptual understanding of cardinality predicts superior school-entry number-system knowledge. *Psychological science*, *29*(2), 191-205.

Gelman, R., & Gallistel, C.R. (1978). The child's understanding of number. Cambridge, Mass.: Harvard University Press.

Gibson, D. J., Gunderson, E. A., Spaepen, E., Levine, S. C., & Goldin-Meadow, S. (2019). Number gestures predict learning of number words. *Developmental science*, *22*(3), e12791.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, *306*(5695), 496-499.

Grabner, R. H., Saalbach, H., & Eckstein, D. (2012). Language-switching costs in bilingual mathematics learning. *Mind, Brain, and Education*, *6*(3), 147-155.

Guillaume, M., & Gevers, W. (2016). Assessing the approximate number system: No relation between numerical comparison and estimation tasks. *Psychological Research, 80*(2), 248-258.

Jara-Ettinger, J., Piantadosi, S., Spelke, E. S., Levy, R., & Gibson, E. (2017). Mastery of the logic of natural numbers is not the result of mastery of counting: Evidence from late counters. *Developmental Science, 20*(6), e12459.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*(4), 498-525.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395-438.

Le Corre, M., Li, P., Huang, B. H., Jia, G., & Carey, S. (2016). Numerical morphology supports early number word learning: Evidence from a comparison of young Mandarin and English learners. *Cognitive psychology*, *88*, 162-186.

Le Corre, M., Van de Walle, G., Brannon, E. M., & Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cognitive psychology*, *52*(2), 130-169.

LeFevre, J. A., Fast, L., Skwarchuk, S. L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. Child development, 81(6), 1753-1767.

LeFevre, J. A., Skwarchuk, S. L., Sowinski, C., & Cankaya, O. (2022). Linking Quantities and Symbols in Early Numeracy Learning. Journal of Numerical Cognition, 8(1), 1-23.

Li, P., Le Corre, M., Shui, R., Jia, G., & Carey, S. (2003). Effects of plural syntax on number word learning: A cross-linguistic study. In *28th Boston University Conference on Language Development*, Boston, MA.

Lin, J. F. L., Imada, T., & Kuhl, P. K. (2012). Mental addition in bilinguals: an fMRI study of task-related and performance-related activation. *Cerebral Cortex*, *22*(8), 1851-1861.

Marchand, E. & Barner, D. (2019). The Acquisition of French Un. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Marchand, E., Lovelett, J. T., Kendro, K., & Barner, D. (2022). Assessing the knower-level framework: How reliable is the Give-a-Number task?. *Cognition*, *222*, 104998.

Marchand, E., Wade, S., Sullivan, J., & Barner, D. (2020). Language-specific numerical estimation in bilingual children. *Journal of Experimental Child Psychology*, *197*, 104860.

Meyer, C., Barbiers, L. C. J., & Weerman, F. (2020). Many systems, one strategy: Acquiring ordinals in Dutch and English. *Glossa: a journal of general linguistics*, *5*(1), 1-31.

Mondt, K., Struys, E., Van den Noort, M., Balériaux, D., Metens, T., Paquier, P., Van de Craen, P., Bosch, P., & Denolin, V. (2011). Neural differences in bilingual children's arithmetic processing depending on language of instruction. *Mind, Brain, and Education*, *5*(2), 79-88.

Negen, J., & Sarnecka, B. W. (2012). Number-concept acquisition and general vocabulary development. *Child development*, *83*(6), 2019-2027.

Nikoloska, A. (2009). Development of the cardinality principle in Macedonian preschool children. *Psihologija*, *42*(4), 459-475.

Piantadosi, S. T., Jara-Ettinger, J., & Gibson, E. (2014). Children's learning of number words in an indigenous farming-foraging group. *Developmental science*, *17*(4), 553-563.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, *306*(5695), 499-503.

Saalbach, H., Eckstein, D., Andri, N., Hobi, R., & Grabner, R. H. (2013). When language of instruction and language of application differ: Cognitive costs of bilingual mathematics learning. *Learning and Instruction*, *26*, 36-44.

Salillas, E., & Wicha, N. Y. (2012). Early learning shapes the memory networks for arithmetic: evidence from brain potentials in bilinguals. Psychological science, *23*(7), 745-755.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*(3), 662-674.

Sarnecka, B. W., Kamenskaya, V. G., Yamana, Y., Ogura, T., & Yudovina, Y. B. (2007). From grammatical number to exact numbers: Early meanings of 'one','two', and 'three'in English, Russian, and Japanese. *Cognitive psychology*, *55*(2), 136-168.

Sarnecka, B. W., & Lee, M. D. (2009). Levels of number knowledge during early childhood. *Journal of experimental child psychology*, *103*(3), 325-337.

Sarnecka, B. W., Negen, J., & Goldman, M. C. (2018). Early number knowledge in dual-language learners from low-SES households. In *Language and culture in mathematical cognition* (pp. 197-227). Academic Press.

Sarnecka, B. W., Negen, J., Scalise, N. R., Goldman, M. C., & Rouder, J. (2021). The real preschoolers of Orange County: Early number learning in a diverse group of children.

Schneider, R. M., Sullivan, J., Marušič, F., Biswas, P., Mišmaš, P., Plesničar, V., & Barner, D. (2020). Do children use language structure to discover the recursive rules of counting? *Cognitive Psychology*, *117*, 101263.

Sella, F., & Lucangeli, D. (2020). The knowledge of the preceding number reveals a mature understanding of the number sequence. *Cognition*, *194*, 104104.

Slusser, E. B., Santiago, R. T., & Barth, H. C. (2013). Developmental change in numerical estimation. *Journal of Experimental Psychology: General*, *142*(1), 193.

Spaepen, E., Coppola, M., Flaherty, M., Spelke, E., & Goldin-Meadow, S. (2013). Generating a lexicon without a language model: Do words for number count?. *Journal of Memory and Language*, *69*(4), 496-505.

Spaepen, E., Gunderson, E. A., Gibson, D., Goldin-Meadow, S., & Levine, S. C. (2018). Meaning before order: Cardinal principle knowledge predicts improvement in understanding the successor principle and exact ordering. *Cognition*, *180*, 59-81.

Spelke, E. S. (2017). Core knowledge, language, and number. *Language Learning and Development*, *13*(2), 147-170.

Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition*, *78*(1), 45-88.

Starkey, P., & Cooper Jr., R. G. (1995). The development of subitizing in young children. *British Journal of Developmental Psychology*, *13*(4), 399-420.

Sullivan, J., Frank, M. C., & Barner, D. (2016). Intensive math training does not affect approximate number acuity: Evidence from a three-year longitudinal curriculum intervention. *Journal of Numerical Cognition, 2*(2).

U.S. Census Bureau (2021). Quick Facts, California. Retrieved from https://www.census.gov/quickfacts/CA.

Van Rinsveld, A., Brunner, M., Landerl, K., Schiltz, C., & Ugen, S. (2015). The relation between language and arithmetic in bilinguals: insights from different stages of language acquisition. *Frontiers in psychology*, *6*, 265.

Venkatraman, V., Siong, S. C., Chee, M. W., & Ansari, D. (2006). Effect of language switching on arithmetic: A bilingual fMRI study. *Journal of Cognitive Neuroscience*, *18*(1), 64-74.

Wagner, K., Kimura, K., Cheung, P., & Barner, D. (2015). Why is number word learning hard? Evidence from bilingual learners. *Cognitive psychology*, *83*, 1-21.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155-193.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*(6389), 749-750.

GENERAL DISCUSSION

The advancement of mathematical thinking through societies rests on the construction of abstract numerical representations that are unique to human cognition. An example of such an achievement is the capacity to represent large exact numbers, which transcends the limits of our approximate nonlinguistic representations of numerosity grounded in perception. While a growing body of studies points to the role of natural language as an important factor in the development of human mathematical capacities (Carey & Barner, 2019; Le Corre & Carey, 2007; Pica et al., 2004; Spaepen et al., 2013; Spelke, 2017), the question of how exactly language contributes to this remains unclear. Over the past decades, a number of studies have used the approach of studying bilinguals to investigate the role of language in numerical development (Spelke & Tsivkin, 2001; Wagner et al., 2015). In particular, examining the transfer and differences in skills across languages can help elucidate the role that language-specific experiences play in the development of numerical representations. In this dissertation, I explored the development of a basic numerical skill, namely, verbal numerical estimation, as a case study to probe how language-specific experiences drive change in number concepts. Specifically, I explored how bilingual 3- to 7-year-old children estimate large and small cardinalities across their two languages.

Exploring the role of language in estimation can enable us to further our understanding of how linguistic knowledge shapes the development of number concepts. In a standard estimation task, participants are asked to make verbal estimates of the cardinality of flashed arrays of dots. Because of the rapid presentation of arrays, participants are prevented from counting and therefore need to rely on their intuitive sense of how number words are represented by nonlinguistic magnitudes. Studying estimation can thus inform us of how our linguistic representations of

number are mapped to nonlinguistic magnitude representations. In this thesis, I presented evidence that individual differences in how children make numerical estimates are not purely due to changes in the acuity of their nonlinguistic representations of magnitudes but that they are also due to changes in language-specific knowledge of their counting system. In particular, language-specific experiences affect the nature of mappings between both small and large number words and their respective non-linguistic representations. In the case of small numbers, I argued that language-specific differences in estimation (i.e., subitizing) were due to differences in the strength of item-based associative mappings between individual words and their respective cardinalities. In the case of large numbers, I presented evidence that language-specific differences in estimation were due to differences in analogical mappings between the structure of the count list and nonverbal representations of magnitudes. In addition, I argued that these effects of language-specific experience emerge as soon as children begin number word learning and persist late into development. Finally, I provided evidence that studies investigating bilingual children should take into consideration reliability issues associated with different tasks in their assessment of differences across bilinguals' languages.

In Chapter 1, I examined whether bilingual French-English children aged 5 to 7 years of age show differences in their verbal numerical estimation of large arrays of dots across their two languages. In this study, participants were presented with flashed arrays of dots and were asked to estimate the number of dots in each of their two languages. I found that estimation accuracy differed across children's two languages, and I provided several pieces of evidence in *post hoc* analysis indicating that those differences were not due to disparities in bilingual children's access to number words across their two languages (e.g., how high they could count). Given that bilinguals had the same nonlinguistic representations across languages, these results support the

view that the differences observed in estimation were due to language-specific differences in children's knowledge of how the number words in their count lists are structured. These results are in line with the hypothesis that estimation abilities rely in large part on Structure Mapping – a global mapping mechanism between the linguistic (i.e., count list) and nonlinguistic (ANS values) systems of magnitude representation founded on an analogy between the structure of these two systems. We also investigated what type of knowledge about the structure of their count lists differed across languages. Specifically, children could either have different knowledge of the order or the distances (or both) of number words in their two count lists. Our results suggest that the differences in estimation observed across languages were due to how they represent the relative distances between number words in their count lists across their two languages, because children provided estimates that were well ordered in both languages. Hence, the findings of Chapter 1 provide novel evidence of differences in language-specific knowledge of the structure of the count list across languages. In particular the relative distances between number words induces differences in the mappings that children establish between number words and perceived magnitudes.

In Chapter 2, I explored a potential methodological issue that can arise when testing the bilingual population, namely, that of test-retest reliability. When testing bilinguals twice across their two languages, it can be challenging to distinguish the amount of variability resulting from measurement reliability inherent to the tasks used versus true differences in knowledge across languages (e.g., in number word knowledge), without knowing what the test-retest reliability of the tasks is. In other words, if a task has poor reliability, differences in knowledge, and therefore in performances, across languages could be expected even if a child were simply tested in the same language twice. To address this question – and thus the interpretation of past work on number word

learning in bilinguals – in Chapter 2, I assessed the test-retest reliability of the Give-a-Number task, a task viewed as the Gold Standard in the field of number word learning. Across three experiments, I presented evidence that although there was an important amount of variability in the reliability of different knower levels (particularly within the group of subset-knowers), the reliability of the titrated and non-titrated versions of administration of Give-N was overall high. This was particularly true for the group of CP-knowers. Overall, the findings of this study allow us to put into perspective the data from previous studies on the acquisition of number words in bilingual and monolingual children and provide reliability indexes for future studies interested in using Give-N in the context of bilingualism or not. These findings also address a growing concern in the field of number cognition about the reliability and validity of the tasks employed. For example, recent studies have questioned the limits on the types of N-knower a child could be (e.g., a 6- or 15-knower) and consequently what it means to be classified as a CP-knower (e.g., Krajcsi et al., 2018). However, our finding of a high reliability for CP-knowers suggests that, although it remains unclear what the entire set of inferences CP-knowers have access to when they reach this stage, at least, they can reliably and successfully deploy counting procedures when constructing sets within their counting list.

In Chapter 3, I investigated when language-specific experiences begin to impose differences in number concepts by exploring bilinguals' ability to perform small number estimation – a process referred to as subitizing. To do so, I relied on the findings of Chapter 2 and addressed this question in CP-knowers, a group of children that show high test-retest reliability. More precisely, in this chapter, I revisited the findings of Wagner et al., (2015) who tested bilinguals and found that children classified as CP-knower in one language were likely to be CP-knower in their second language as well. These results might lead to the conclusion that once

children reach the CP-knower stage in one language, the differences in mappings between number words and cardinality found across languages (i.e., in the case of subset-knowers) disappears. In Chapter 3, I challenge this assumption by testing the subitizing abilities of bilingual CP-knowers across their two languages. Specifically, I examined whether bilingual Spanish-English and German-English children aged 3- to 6-year-old make different verbal numerical estimates of small arrays of dots across their two languages. Against the assumption raised by the findings of Wagner et al., in this chapter I presented evidence that bilingual CP-knowers still make more accurate verbal estimates of small sets in their dominant number language, that is, the language in which they could count the highest. This finding also expanded the results of Chapter 1 by showing that the differences in mappings across languages for large numbers is also present for small numbers and with a younger population of bilinguals. Finally, the findings of Chapter 3 provide support for the view that the counting procedure (i.e., the knowledge CP-knowers have) could be learned in parallel to the acquisition of associative mappings between small number words and cardinalities rather than the result of it, as proposed by a bootstrapping account of number word learning (Carey, 2004).

Taken together, the findings of Chapter 1 and 3 provide evidence that language specific experiences influence the mappings between the linguistic and nonlinguistic systems of numerical representations, for both small and large numbers. However, the data presented in this thesis raise questions about other types of knowledge that might be dependent upon language specific experience. For example, past studies have shown that some types of knowledge such as math facts (e.g., 2 x 2 = 4) do not transfer across languages. Here, we added to this literature by showing that some knowledge about the structure of the count list does not transfer across languages and associative mappings even when children learn how to count accurately across languages. This

raises a question: do other forms of numerical knowledge transfer across languages? For example, intuition about the productivity of a counting system? Previous studies have shown that the morphological structure of a count system can have an impact on children's ability to extract the rules of their counting system (e.g., the rule that to form decade labels in English, one must add "ty") and consequently their ability to keep counting when prompted to (Schneider et al., 2020). Bilingual children who are navigating two very different counting systems might be more limited in one system compared to the other because of a disparity in their language specific experience with each count list.

One central conceptual component of this thesis concerns the type of knowledge and basic numerical skills that transfer across languages when children are classified as CP-knowers. Studying bilinguals can inform us of the sources of variability in monolingual children's understanding of numerical concepts. Individual differences in monolinguals could emerge from children relying on different types of representations (e.g., ANS values, associative mappings between small number words and cardinality, counting) when performing numerical tasks. The question of what drives CP-knowers to rely more on one source of knowledge (or representation) compared to another is still a matter of debate in the literature. This could be addressed in bilingual children who have different knowledge and experience across languages. This direction of research has the potential to further elucidate the sources of individual differences in basic math skills in children.

Overall, the findings in this dissertation identify language-specific experiences as an important source of change in number conceptual development. While several questions remain to

be explored, the work in this dissertation provides an additional piece of evidence on the importance of how language and symbolic representations come to explain perception.

References

Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, *133*(1), 59-68.

Carey, S., & Barner, D. (2019). Ontogenetic origins of human integer representations. *Trends in cognitive sciences*, *23*(10), 823-835.

Krajcsi, A., Fintor, E., & Hodossy, L. (2018). A refined description of preschoolers' initial symbolic number learning.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395-438.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, *306*(5695), 499-503.

Schneider, R. M., Sullivan, J., Marušič, F., Biswas, P., Mišmaš, P., Plesničar, V., & Barner, D. (2020). Do children use language structure to discover the recursive rules of counting?. *Cognitive psychology*, *117*, 101263.

Spaepen, E., Coppola, M., Flaherty, M., Spelke, E., & Goldin-Meadow, S. (2013). Generating a lexicon without a language model: Do words for number count?. *Journal of Memory and Language*, *69*(4), 496-505.

Spelke, E. S. (2017). Core knowledge, language, and number. *Language Learning and Development*, *13*(2), 147-170.

Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition*, *78*(1), 45-88.

Wagner, K., Kimura, K., Cheung, P., & Barner, D. (2015). Why is number word learning hard? Evidence from bilingual learners. *Cognitive psychology*, *83*, 1-21.